The Art of Mathematics Retrieval

Petr Sojka et al.

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

Informatics Colloquium, FI MU, Brno, Czech Republic
November 8th, 2011

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Why Math Retrieval (TEX math search)?

Searching is crucial part of *accessibility* of the great stuff you all create, usually with the lot of *mathematics* with formulae and equations.

How to pose questions about mathematics?

Similarity as in MUFIN (pictures), Sketch Engine (text attributes)?

Math in TEX notation?

- Compact and logical expression of formulae, quickest entering of them into a query or a document.
- A picture is worth a thousand words, “a mathematical formulae is worth of hundred words” (Ross Moore).

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

- Because of G? (G as in Google, Globalization,…).
- The *vast* treasure of mathematical papers; 140,000 new papers in Zentralblatt MATH expected this year, most of them authored in TEX math notation. All mathematics ever publisher is estimated at 100,000,000 pages (3,500,000 articles).
- Search – crucial part (access to data); search is a *gate* to this knowledge; Digital Mathematics Library (DML) without math-aware search is an oxymoron.
- Text and keyword based search? No problem (Google, review databases); *success*.
- Mathematics formulae search? It *is* a problem (either in Google or in the review databases); more or less a *failure so far*.

Motivation for MSE (DML panel discussion)

Q: “What functionality and incentives would made a working mathematician to login and use a modern DML as EuDML?”

A: **“Math formulae search.”**

Prof. James Davenport, CEIC member, MKM 2011 PC chair, on panel at DML 2011 workshop in Bertinoro as a reply

Motivation for using a MSE (including formulae) – cont.

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.

Compare `google://Einstein` with math-aware search of "`Einstein $E=mc^2$`" over arXiv.

Motivation for using a MSE (search examples) – cont.

- Search problem formulation: given query containing text and formulae, find the most relevant documents.
- 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 application supports math formulae search.

Take-off message from this talk

Yes, you can! (in our MIaS system)

The rest of the talk: how is it actually done*, how are the formulae indexed and how the search is performed to be useable on DL with hundreds of millions formulae?*

Towards math search engine (MSE) – existing players

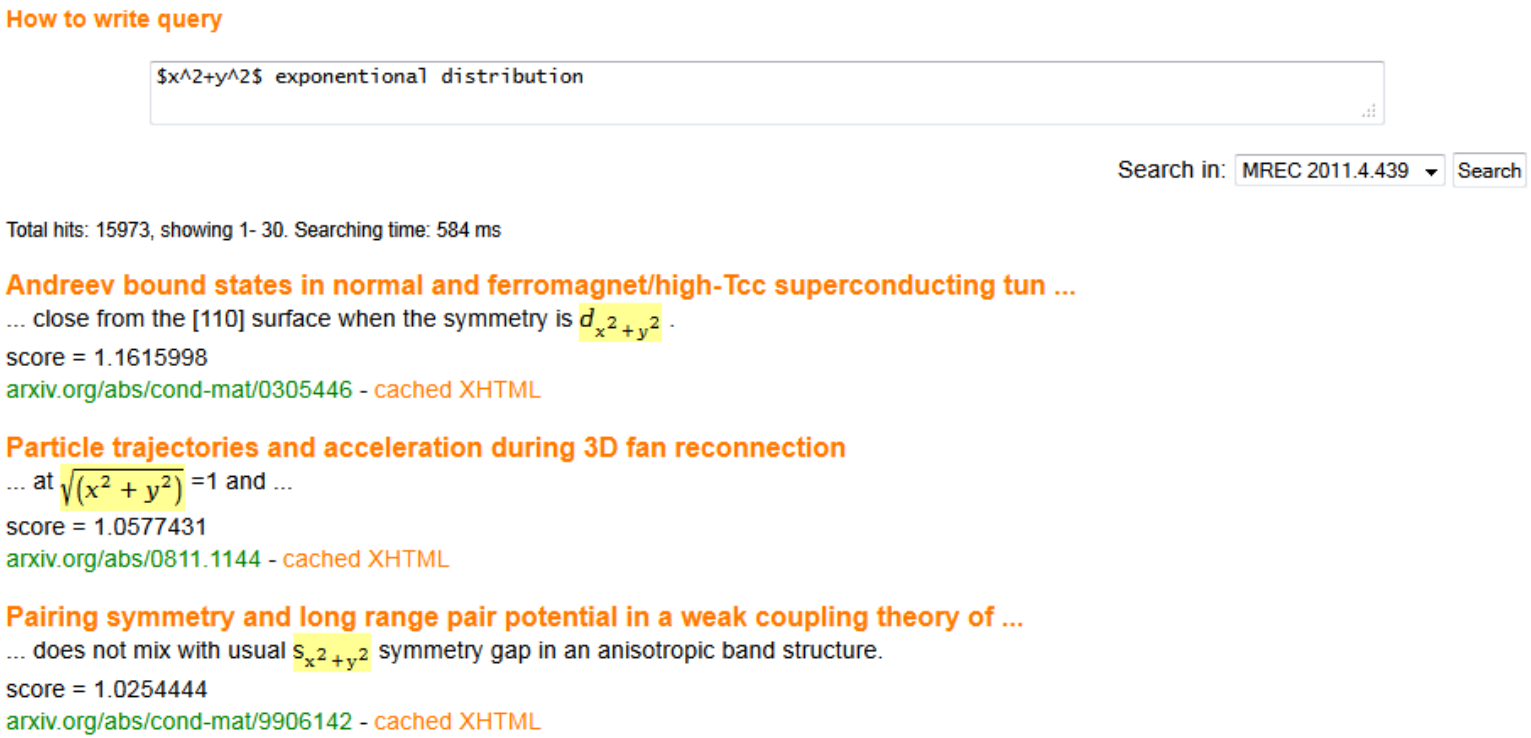
- Niche market for big players (as Google), attempts to solve by publishers (LaTeXSearch by Springer).
- Many challenges: heterogenity of math representation, notation, semantics handling, no established and accepted user interface and query language.
- Numerous attempts to solve the problem: MathDex, EgoMath, LATEXSearch, LeActiveMath, DLMF equation search, MathWebSearch, but none accepted by the community as *the* MSE.

Existing systems—pros and cons

- **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
- **MathDex**: formerly MathFind * seven digit figure NSF grant by Design Science (Robert Miner) * Lucene based, indexing n-grams of presentation MathML * pioneering conversion effort
- **LaTeXSearch**: MSE offered by Springer * closed source * only for LaTeX math string approximate match based on strings * no formulae structure matching * small database: 3 M formulae from 'random' sources (cf. 200 M in arXiv)
- **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 ask and how to index –dual world of TEX and MathML

Math for *people*: TEX notation wins and is used by people (mostly AMSLATEX fits most needs): $\rightarrow$ TEX notation for querying.

Math for *software* applications: MathML wins and is used by most computer algebra systems, browsers, in workflow of DTP systems: $\rightarrow$ MathML for indexing.

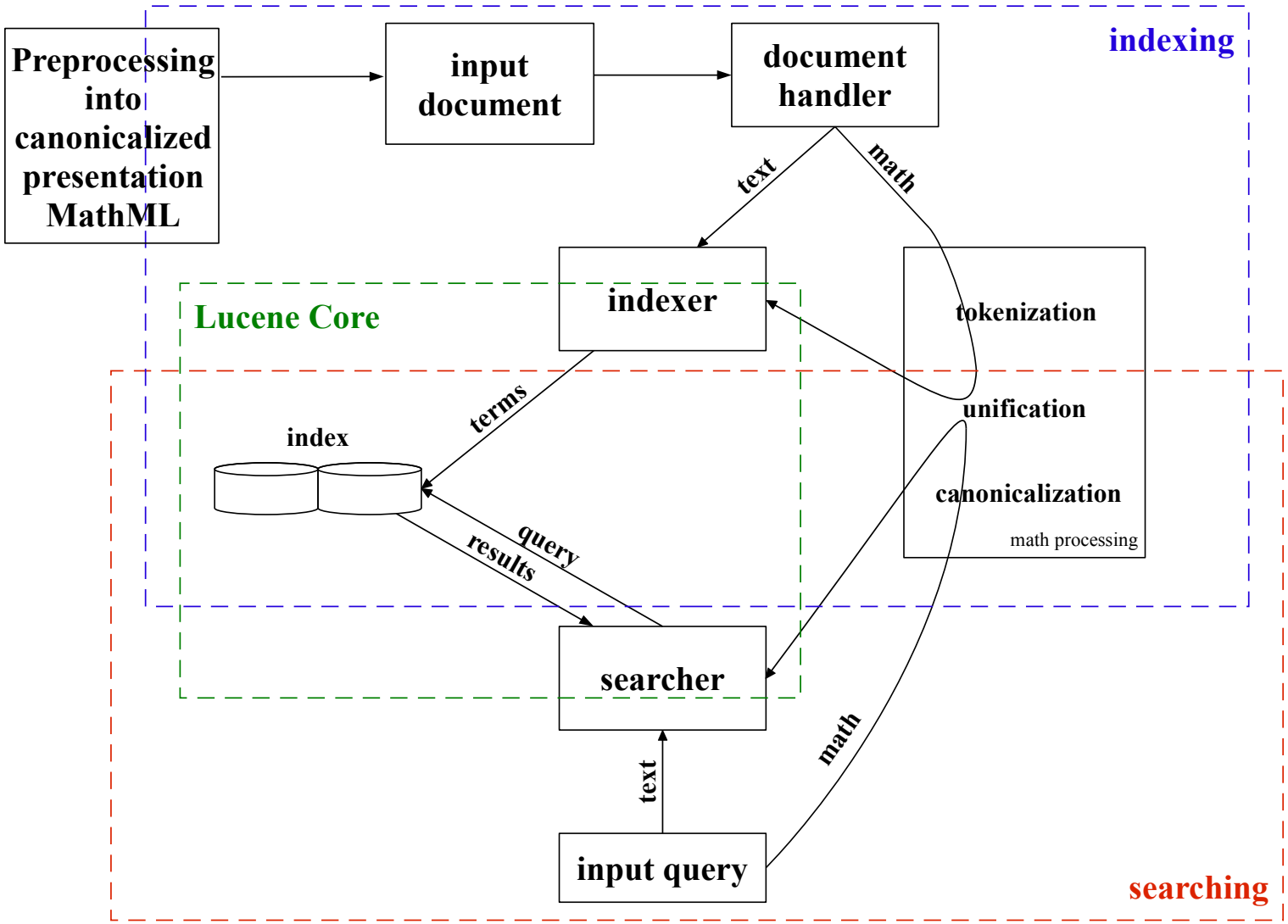
Dual world of querying and indexing languages

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

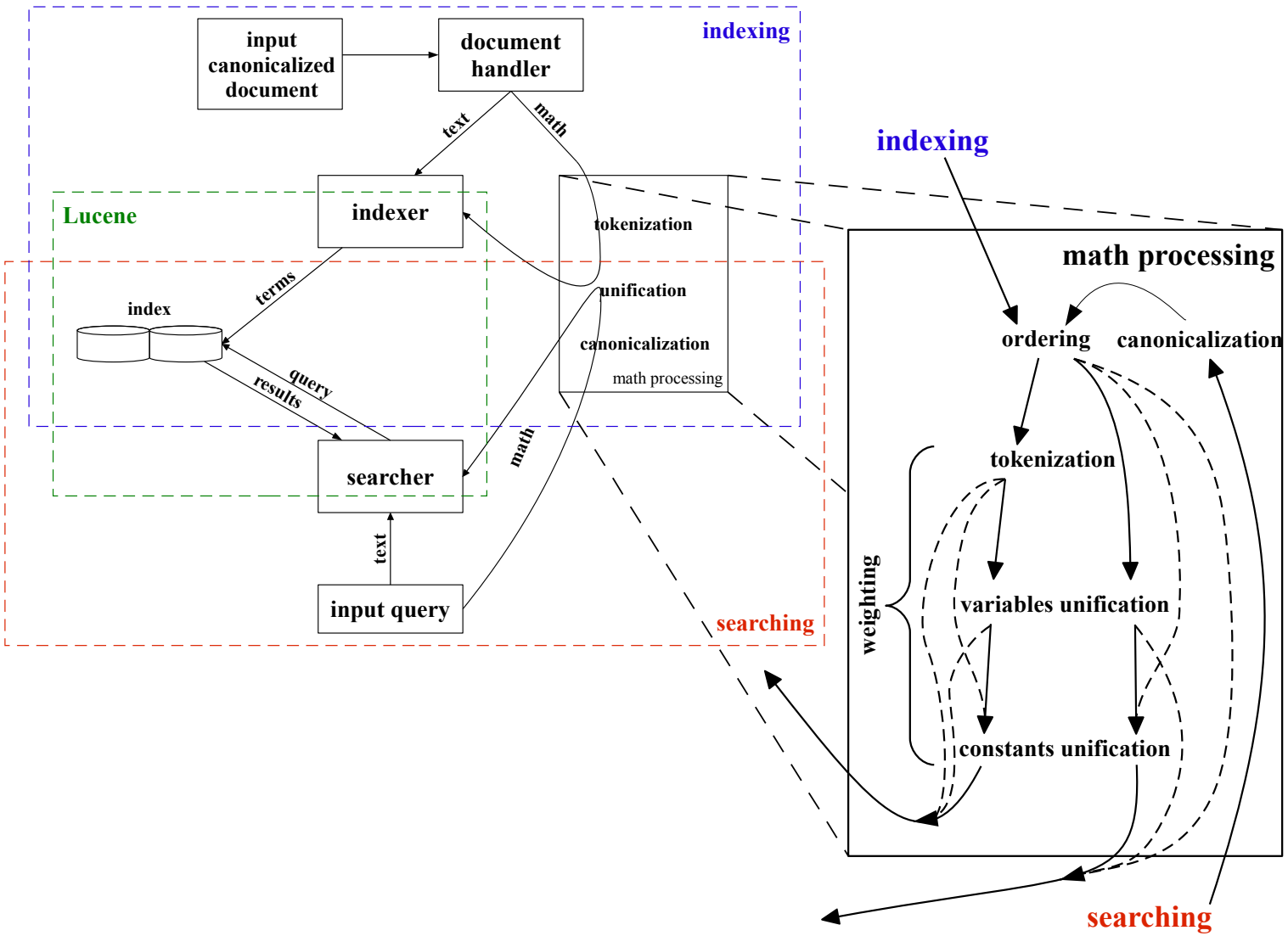
TEXbook's Concert invitation example: there is a name of Czech composer of a song in the index that even does not appear in the invitation.

From text to math: the same idea explored for math (e.g. having multiple representations of a formula (with different 'near synonyms') put in the index).

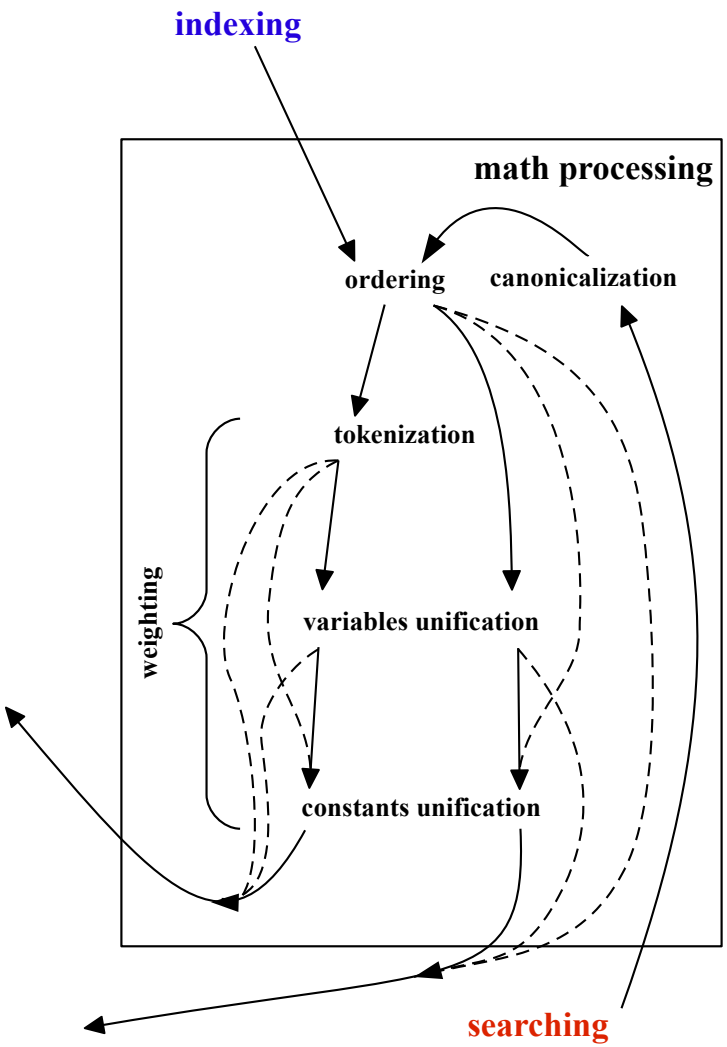
MSE overall design



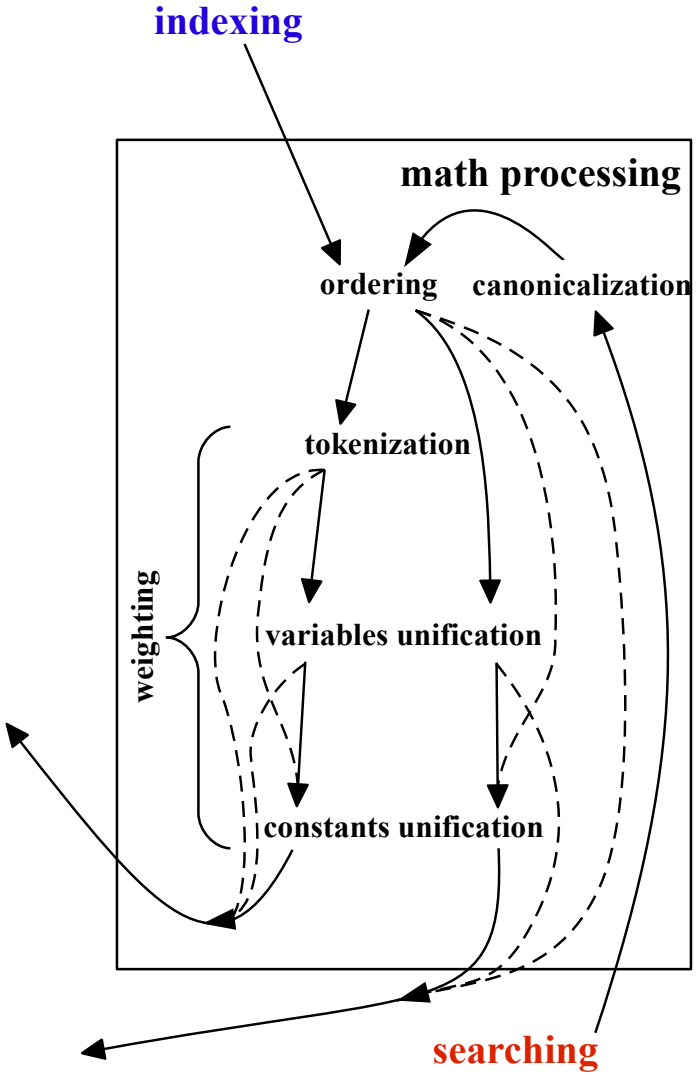
Math indexing design

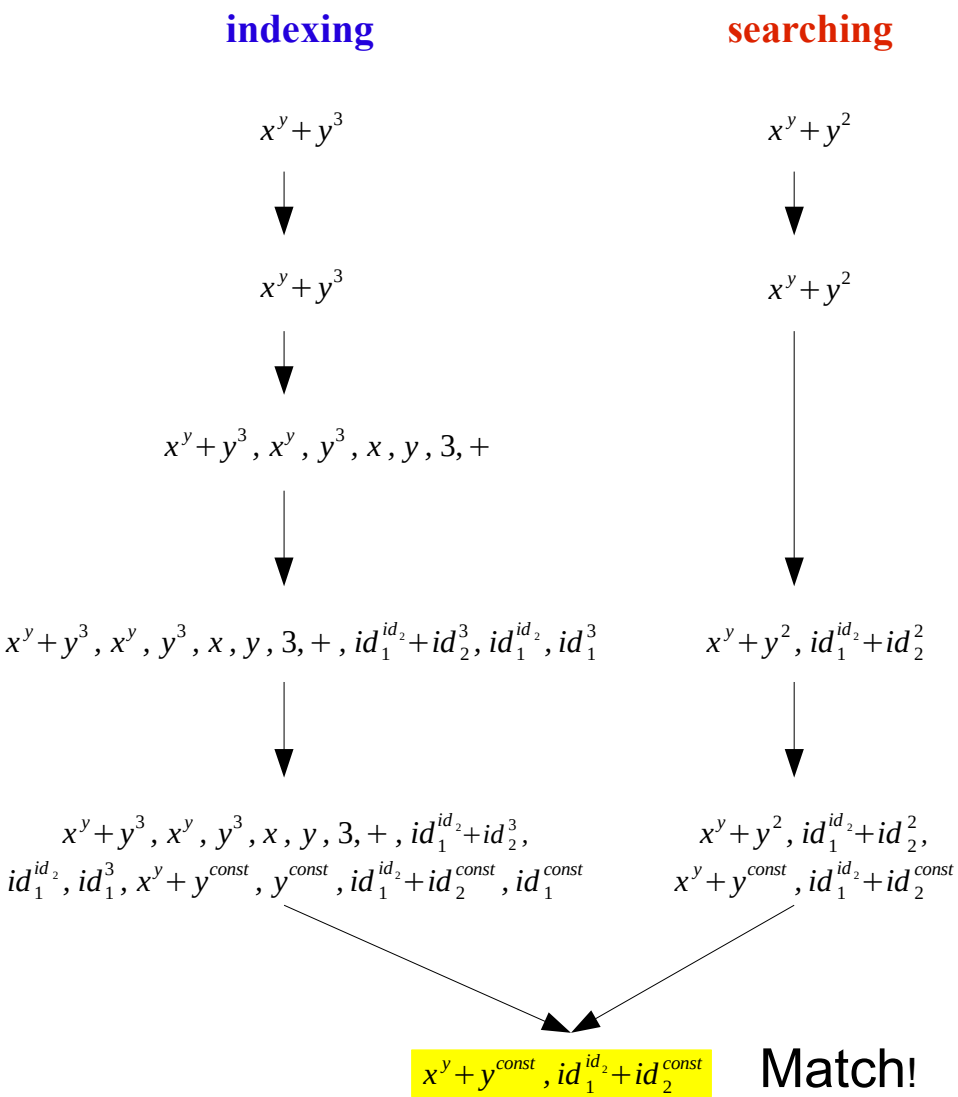


Math formulae indexing processing



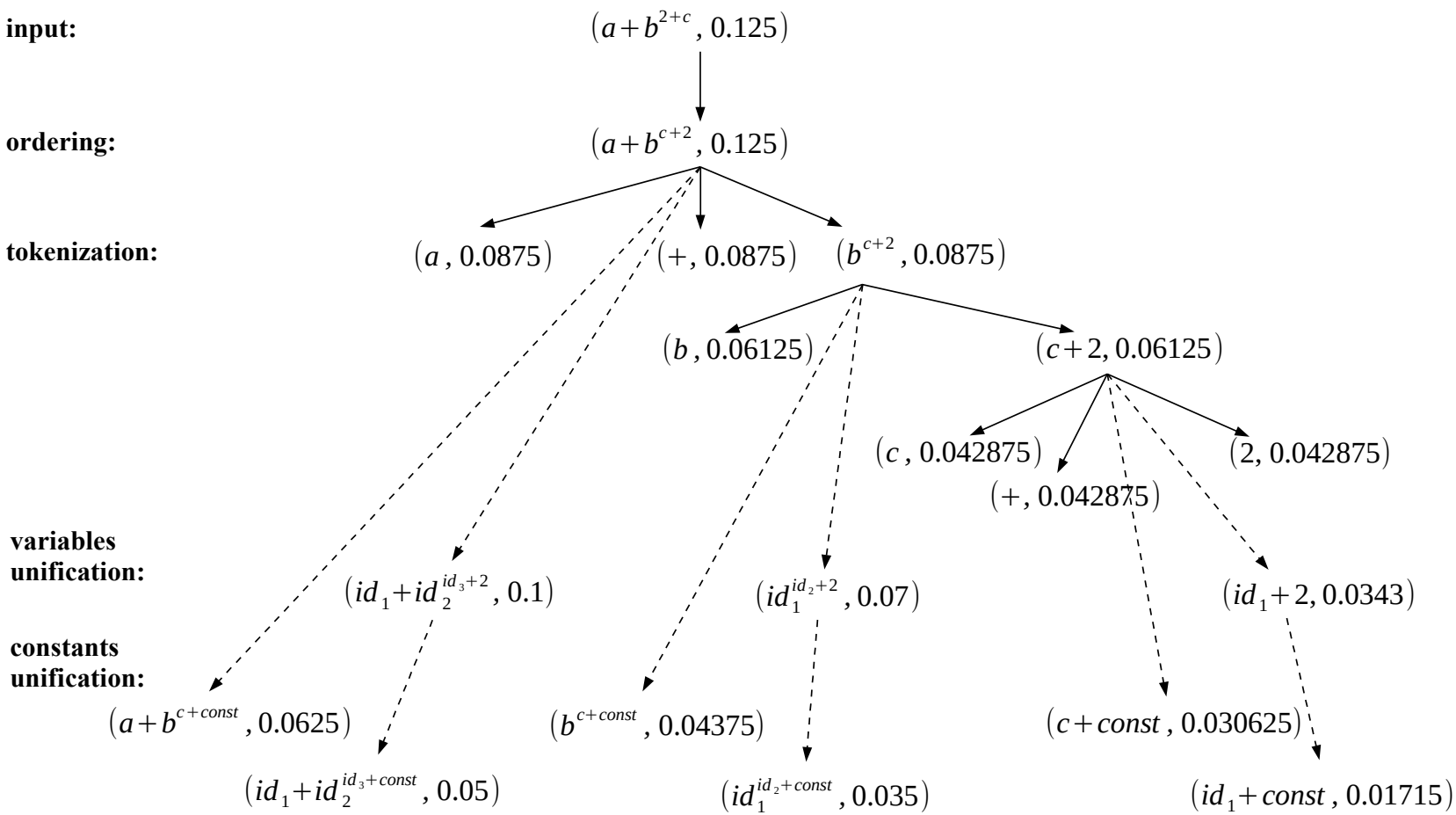
Example





Match!

Formula processing example – subformulae weighting



Weighting

- We used a weighting utility.
- Indexing:
  - initial weight of whole formula $= \frac{1}{number\_of\_nodes}$
  - tokenization – level coefficient $l = 0.7$
  - variables unification – coefficient $v = 0.8$
  - number constants unification – coefficient $c = 0.5$
  - matching `mathvariant` font (under implementation)
- Searching:
  - $result * number\_of\_query\_nodes$

Under implementation: thresholds computed from LSA representations of indexed math terms (by gensim).

Implementation

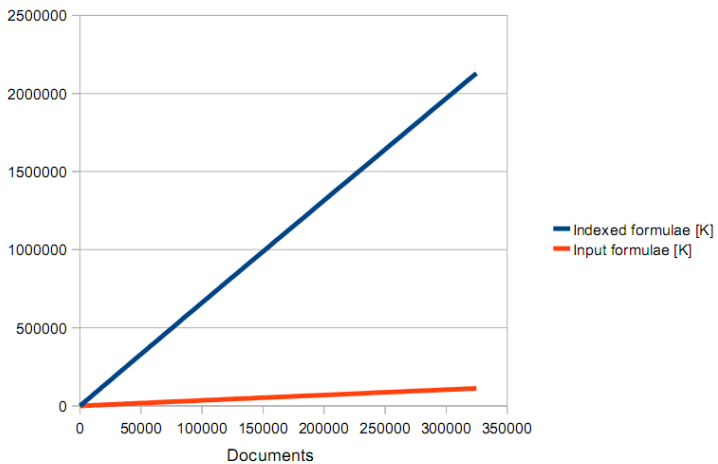
- Java.
- Lucene 3.1.0.
- Mathematical part implements Lucene's interface Tokenizer – able to integrate to any Lucene based system.
- MIaS4Solr plugin was created for the use in Solr in EuDML.
- Textual content – processed by StandardAnalyzer.

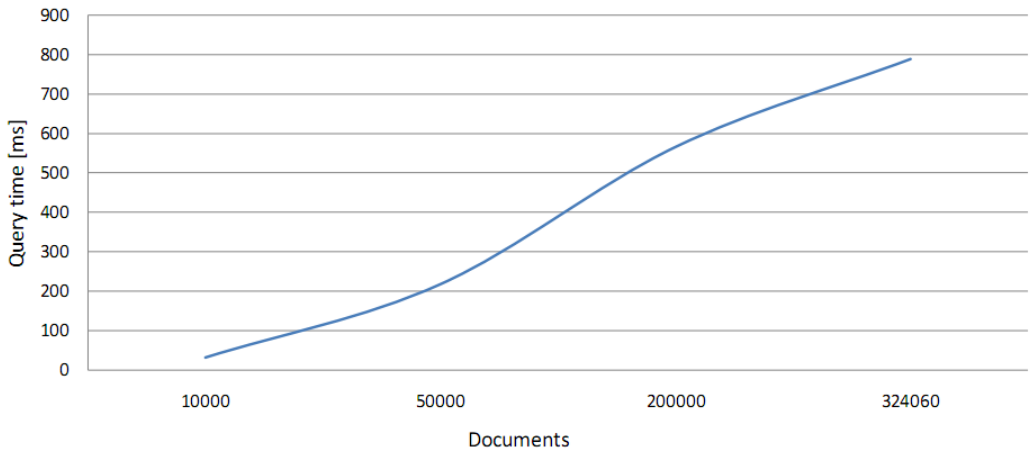
Data used for evaluation: MREC corpus

- 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/.

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.





Formulae search demonstration comments

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

- MathML/TEX input (Tralics [2] for conversion to MathML [7]).
- Canonicalization of the query – UMCL library [1].
- Matched document snippet generation.
- MathJax for nicer math rendering and better portability.

MIaS already integrated in the 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 new 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].

Related work

Work motivated by projects of The European Digital Mathematics Library (EuDML) and The Digital Mathematics Library Czech Republic (DML-CZ).

Related topics researched at FI as part of projects above in LEMMA and NLP laboratories:

- gensim package (*topic modelling for humans*) by Radim Řehůřek.
- pdfRecompressor (*JBIG2 compression enhancements by OCR*,…) by Radim Hatlapatka.
- *TEX to MathML conversion* (Tralics), by Michal Růžička.
- *MathML preprocessing* (normalization and canonicalization) by Michal Růžička, Peter Mravec.

Related work (cont.)

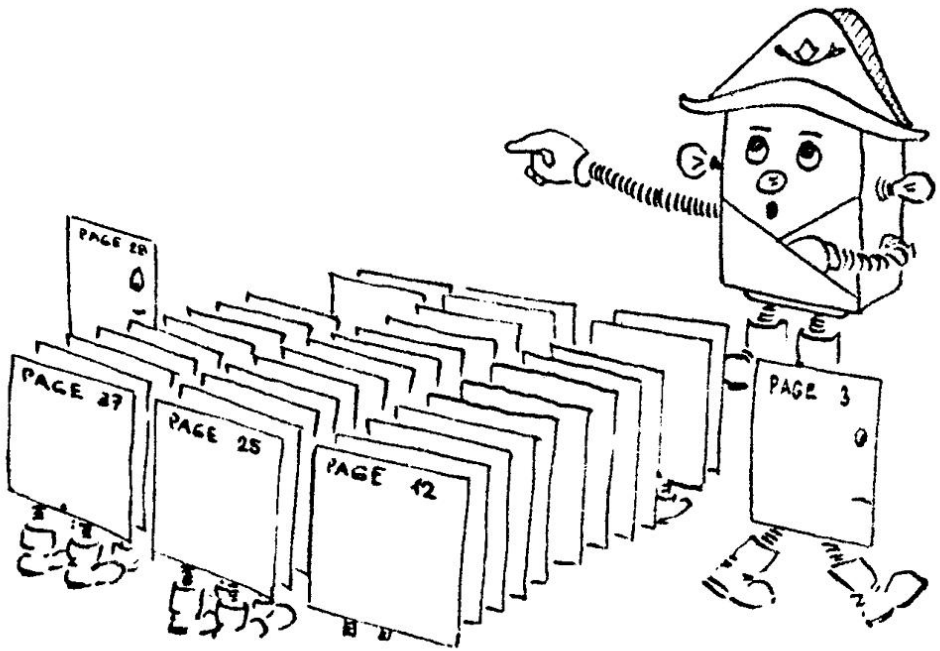
- *Metadata Editor* tool development, metadata enhancements by Petr Kovář, Mirek Bartošek, Vlastimil Krejčíř, Martin Šárfy.
- *(Math) OCR* by Masakazu Suzuki, Radovan Panák, Tomáš Mudrák, Radim Hatlapatka.
- (Meta)data *vizualization* (Visual Browser) by Zuzana Nevěřilová.
- Czech Braille driver with math support by Martin Jarmar.
- And a lot more…

Acknowledgments

- EuDML and DML-CZ project funding.
- Martin Líška (search implementation).
- Michal Růžička, Radim Hatlapatka, Zuzana Nevěřilová, Martin Jarmar, Petr Mravec, Radovan Panák, Tomáš Mudrák, Vítězslav Dostál, Martin Kacvinský.
- Mirek Bartošek, Petr Kovář, Vlastimil Krejčíř, Martin Šárfy.
- Infty group (led by Masakazu Suzuki).
- Numerous authors and contributors of several (mostly OSS) tools used.
- Numerous people discussing and supporting our work.

Questions?

Thank you for your attention.




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/>


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 (July 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 (July 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 (eds.) *Proceedings of DML 2011*, pp. 77–84, Bertinoro, Italy, July 2011. Masaryk University. <http://dml.cz/dmlcz/702604>.