Building Corpora of Technical Texts: Approaches and Tools

Petr Sojka, Martin Líška, Michal Růžička

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

RASLAN 2011, Karlova Studánka, Czech Republic
December 3rd, 2011

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Why STEM corpora and NLP?

- large (e.g. web-scale) corpora such as those created by Google (Google Books Corpus, Google Scholar) or by the Sketch Engine (TenTen Corpora) allow a new quality level to solve such tasks as more relevant information retrieval, document clustering, classification and similarity, thesauri and ontology building, better word sense disambiguation, machine translation and many others.
- minority languages or domain specifics—language of mathematics—typical in Science, Technology, Engineering, and Mathematics (STEM) *neglected*: no rich lobists and wide user's demand, no mainstream tools support this niche market of 'the Queen of sciences'.

STEM corpora specifics

- *mathematics* with formulae and equations.
- A picture is worth a thousand words (proverb), “a mathematical formulae is worth of hundred words” (Ross Moore).

- “Word and image are one” (Hugo Ball) vs.
- “Word and formulae are one” (Petr Sojka)

Challenges

- “The limits of my language means the limits of my world.” (L. Wittgenstein)
- complete new support for mathematical formulae is needed in corpora handling workflow from its beginning—tokenization; support to handle *rich structures* (e.g. formulae trees).
- establishment of G math representation (G as in Google, Globalization,…) to allow for global methods.
- ambiguity of notation: numerous ways of notating the same mathematical object, that has evolved in some geographical location or language: a *binomial coefficient*:

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n!}{r!(n-r)!} = {}_nC_r = {}^nC_r = C(n,r)$$

- math search – crucial math corpora tool; search is a *gate* to this knowledge; corpora without math-aware search is an oxymoron.

Motivation to tackle these challenges

- DML-CZ project
- EuDML project
- Centre (LC536 topic of research)
- establishing new research area of math NLP

Words and formulae

formulae in queries help to *disambiguate and narrow* search:

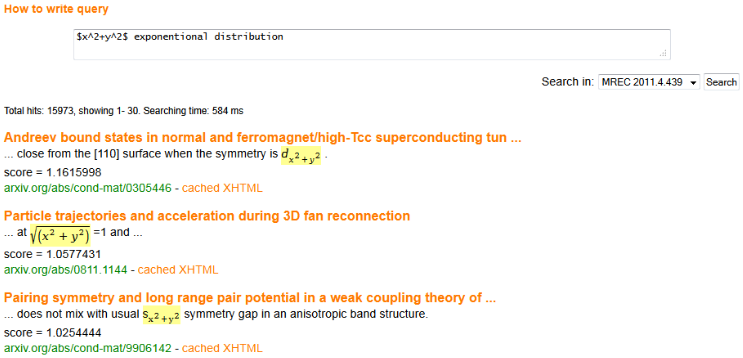
Compare `google://Einstein` with math-aware search of “`Einstein $E=mc^2$`” over arXiv.

Formulae for disambiguation (cont.)

- 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).
- Example 3: physicist looking for theorems about solitons, but mathematicians use these terms for something completely different from my perspective and I do not know how they call those I'm interested in. Putting the equation my solitons are solutions of might be the only way to locate relevant literature.

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

```
$x^2+y^2$ exponentional 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

Math representations – TEX, MathML and M-terms

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

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

Math for *corpora* (indexing and bag of words representation): *M-terms*

Examples of representation

TEX: `$a^2+b$`

MathML:

```
<math>
  <mrow>
    <msup><mi>a</mi><mn>2</mn></msup>
    <mo>+</mo>
    <mi>b</mi>
  </mrow>
</math>
```

M-terms

M-terms for $a^2 + b$:

```
(mi(a),0.08166666),
(mn(2),0.08166666),
(msup(mi(a)mn(2)),0.11666667),
(mo(+),0.11666667),
(mi(b),0.11666667),
(mrow(mi(b)mo(+)msup(mi(a)mn(2))),0.16666667),
(msup(mi(1)mn(2)),0.093333334),
(mrow(mi(1)mo(+)msup(mi(2)mn(2))),0.13333334),
(msup(mi(a)mn(¶)),0.058333334),
(mrow(mi(b)mo(+)msup(mi(a)mn(¶))),0.083333336),
(msup(mi(1)mn(¶)),0.046666667),
(mrow(mi(1)mo(+)msup(mi(2)mn(¶))),0.06666667)
```

M-term compactification

`mrow(msup(mi(a)mn(2))mo(+)mi(b))` is further compacted to `R(J(I(a)N(2))O(+)I(b))` based on a custom tag name dictionary, where `mrow=R`; `msup=J`; `mi=I`; `mn=N` and `mo=O`.

RESTful web service

...mias4gensim/mathprocess?mterm=<math><mrow><mi>a</mi><mo>+</mo><mi>b</mi></mrow></math>

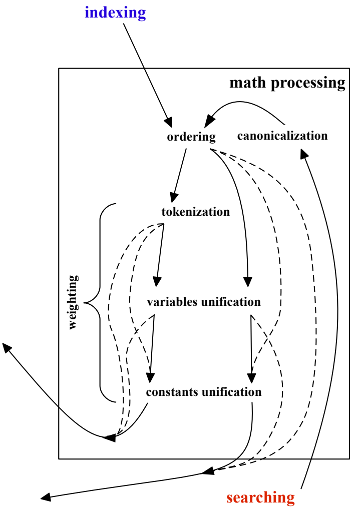
Dual world of querying and indexing languages

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

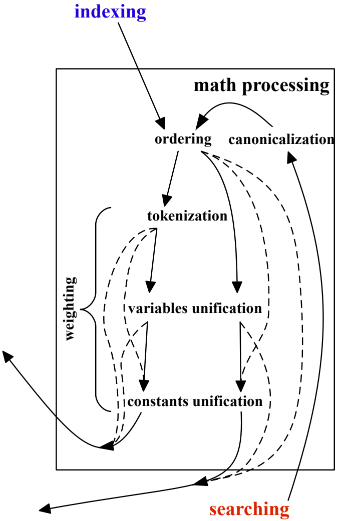
TEXbook's Concert invitation example: there is a name of Czech composer of a song in the index that even does not appear in the invitation.

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

Math formulae indexing processing



Example

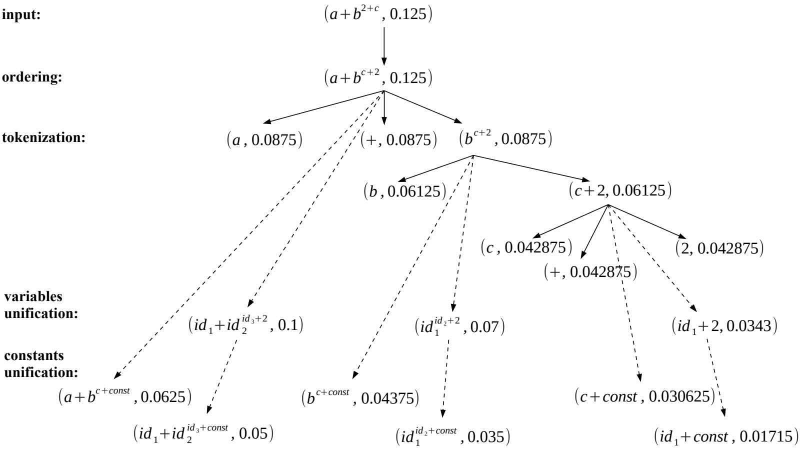


indexing — searching

$$\begin{tikzcd}
x^y+y^3 \arrow[d] & & x^y+y^2 \arrow[d] \\
x^y+y^3 \arrow[d] & & x^y+y^2 \arrow[dd] \\
x^y+y^3, x^y, y^3, x, y, 3, + \arrow[d] & & \\
x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3, id_1^{id_2}, id_1^3 \arrow[d] & & x^y+y^2, id_1^{id_2}+id_2^2 \arrow[d] \\
x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3, id_1^{id_2}, id_1^3, x^y+y^{const}, y^{const}, id_1^{id_2}+id_2^{const}, id_1^{const} \arrow[dr] & & x^y+y^2, id_1^{id_2}+id_2^2, x^y+y^{const}, id_1^{id_2}+id_2^{const} \arrow[dl] \\
 & x^y+y^{const}, id_1^{id_2}+id_2^{const} &
\end{tikzcd}$$

Match!

Formula processing example – subformulae weighting



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

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

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

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

$(c, 0.042875)$ $(+, 0.042875)$ $(2, 0.042875)$

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

constants unification: $(a+b^{c+const}, 0.0625)$ $(id_1+id_2^{id_3+const}, 0.05)$ $(b^{c+const}, 0.04375)$ $(id_1^{id_2+const}, 0.035)$ $(c+const, 0.030625)$ $(id_1+const, 0.01715)$

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).

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/.

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

- First math corpora built, and new representation for math formulae handling designed (M-terms)
- MREC and MIaS project pages: http://nlp.fi.muni.cz/projekty/eudml/mias/

Future work

- Gensim using M-terms
- LDA-frames to disambiguate M-terms
- 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?

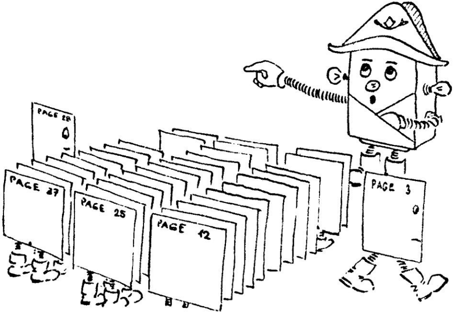
Summary

Corpora for STEM domain need special support and tools.

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

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>.