Towards a Meaning-Aware Math Information Retrieval

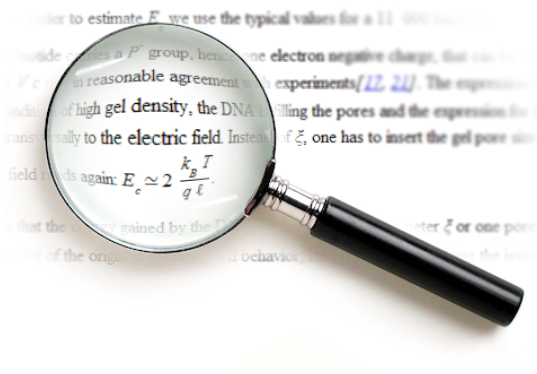
Petr Sojka et al

Masaryk University, Faculty of Informatics, Brno, Czech Republic
<https://mir.fi.muni.cz/>

Wikipedia Subtask Pre-Conference Meeting, NTCIR-11 2014, Tokyo, Japan
December 8th, 2014, 2PM

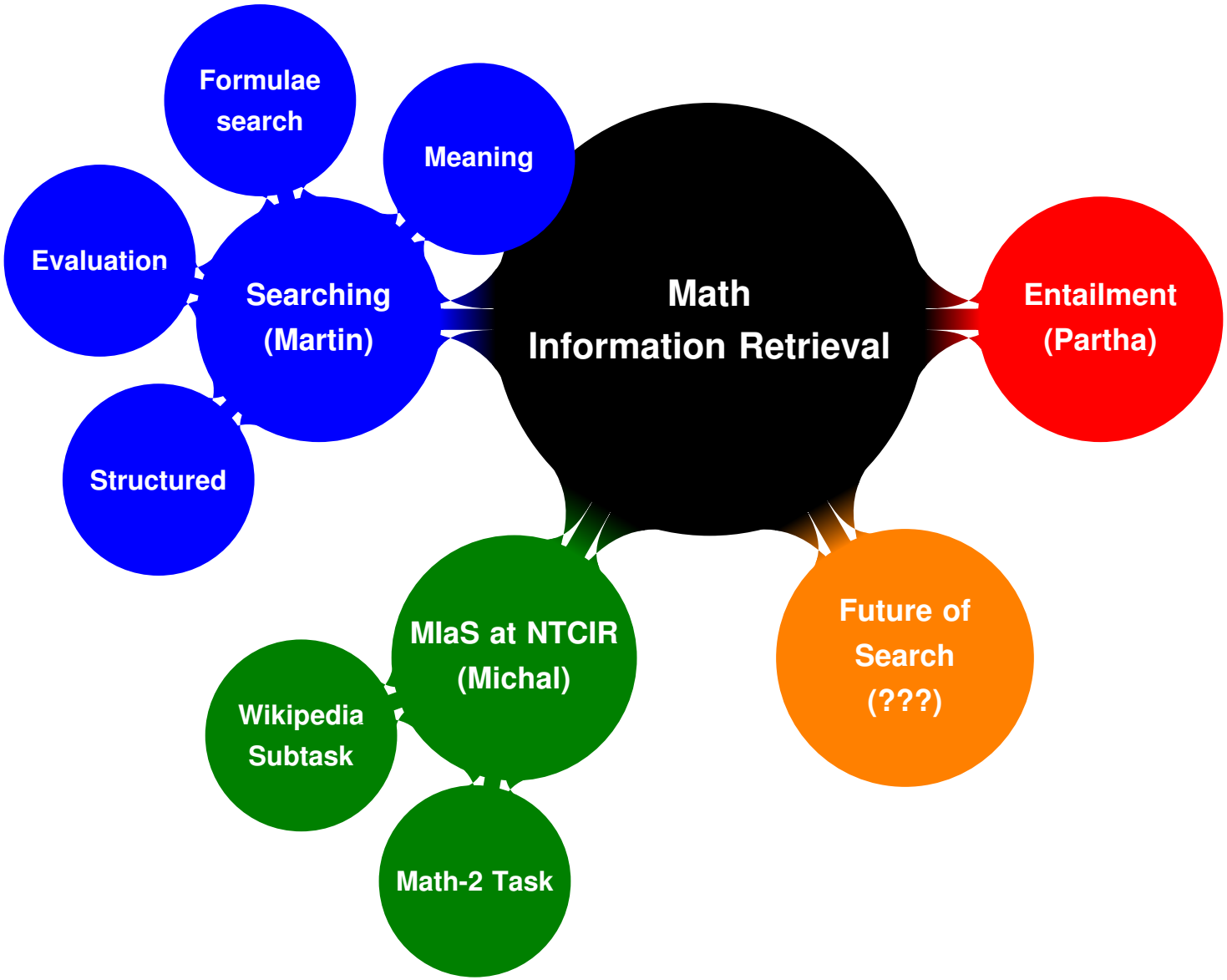
EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA •



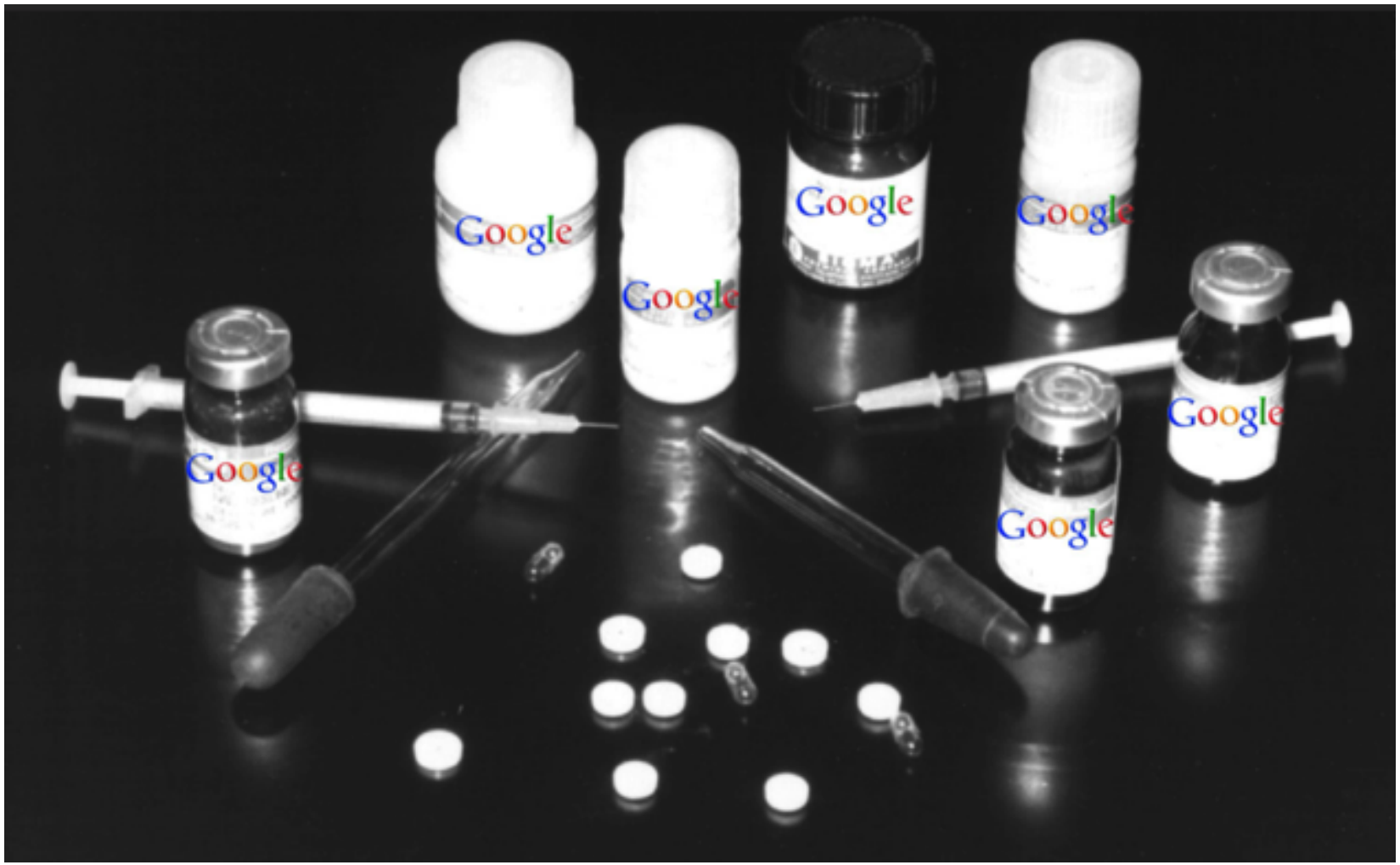
Illustrations by Jiří Franek.

Talk Topics and Take-home Message



Outline

Dependency on Information Retrieval: Information Society Now!



Search: A Gate to Knowledge

Querying and *search*ing *similar* structures more and more important: striving to find the right meaning.

Meaning (semantics) is usually structured compositionally.

Structures: math formulae, syntactic or sentence dependency trees, compositional named entity terms, knowledge base terms.

<http://google.cz/search?q=Kovacik+Rakosnik>

```
$L^{p(x)}$
```

```
https://www.google.cz/search?q="L^{p(x)}"
```

+ without quotes or figures :-).

Nature 454, 263 (2008) | doi:10.1038/454263b

Starting small but adding up: a free maths archive

A small group of researchers is meeting in Birmingham, UK, later this month to plan a free digital library of mathematics.

All the mathematical literature ever published runs to more han 50 million pages, with around 75,000 articles added each year. Over the past decade there have been several attempts to make this prodigious body of work accessible in a single digital archive, but so far none has succeeded.

A group of mathematicians intends to change this. They have started small, with a handful of digitization projects in Poland, Russia, Serbia and the Czech Republic. In a few years they hope to unite these repositories with their western European counterparts in an archive to be hosted by the European Union, according to the organizer, Petr Sojka, an informatics scientist at Masaryk University in Brno in the Czech Republic. Eventually this pan-European archive could be expanded globally, he says.

To make such an archive easier to search, researchers have found ways to guess the subject of a paper on the basis of the frequency of symbols in it. But there will be many more-practical challenges, such as finding the funds to scan millions of old papers and striking deals with publishers who hold rights to them.

It may already be too late to build a single free mathematical archive, according to John Ewing, head of the American Mathematical Society, which maintains a list of more than 1,500 journals whose archives have already been digitized. "A few years ago, this model had the potential to change the mathematics journal literature in profound ways," he says. But most publishers have rushed to scan their own archives in order to lock them up and sell them to libraries.

"While the effort to digitize the smaller collections is admirable, and it's certainly worthwhile, it's unlikely to effect a larger change," says Ewing. ■

Jascha Hoffman

Workshop series *Towards a Digital Mathematics Library* founded to tackle numerous challenges identified during DML-CZ project.

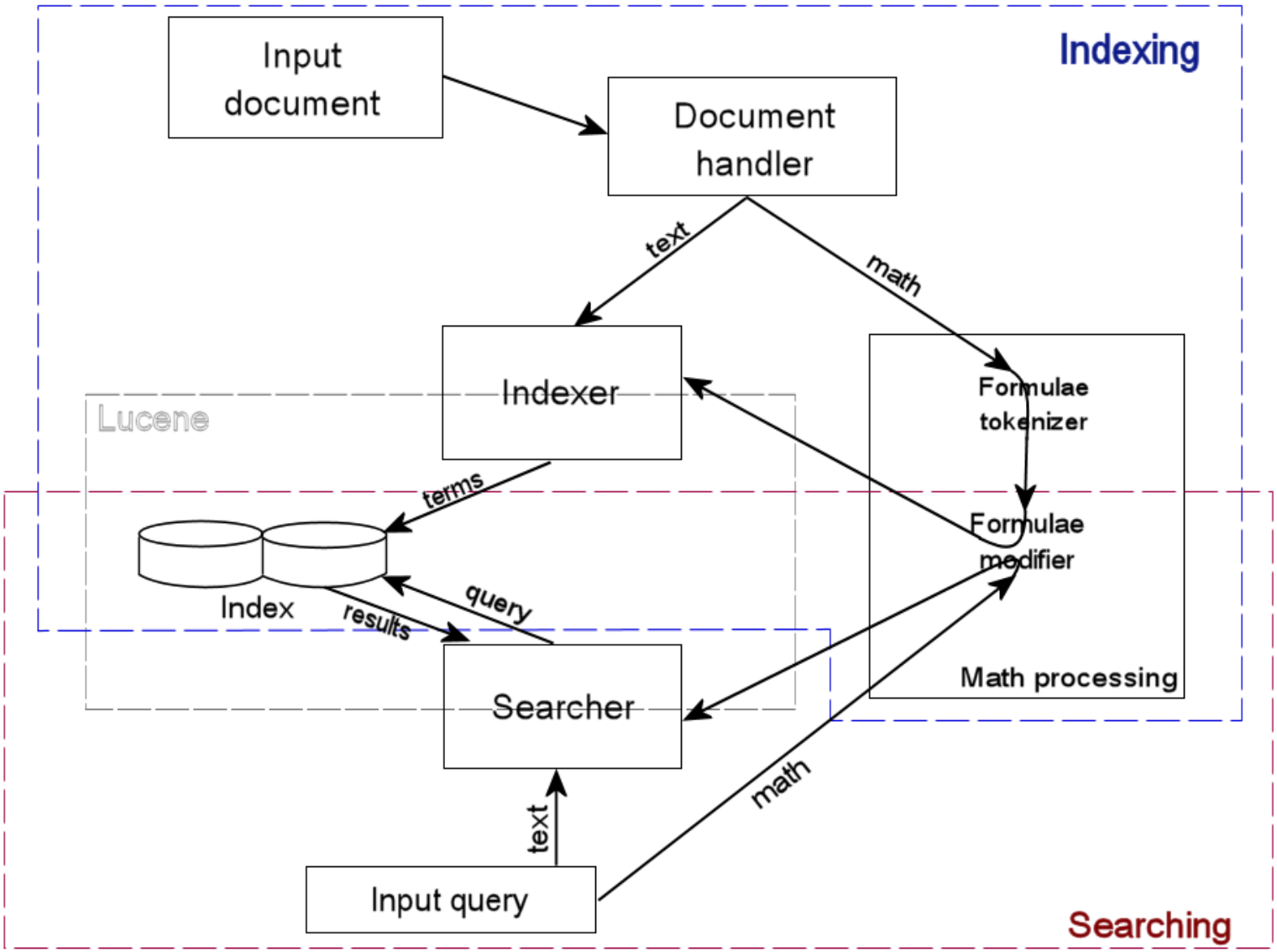
Math-aware Search and Indexing

- Conventional searching approaches are not applicable for math.
- Usage of existing mathematical search engines (MathDex, EgoMath, LaTeXSearch, LeActiveMath, MathWebSearch) problematic.
- New Math Indexer and Searcher (MIaS) developed at MU.

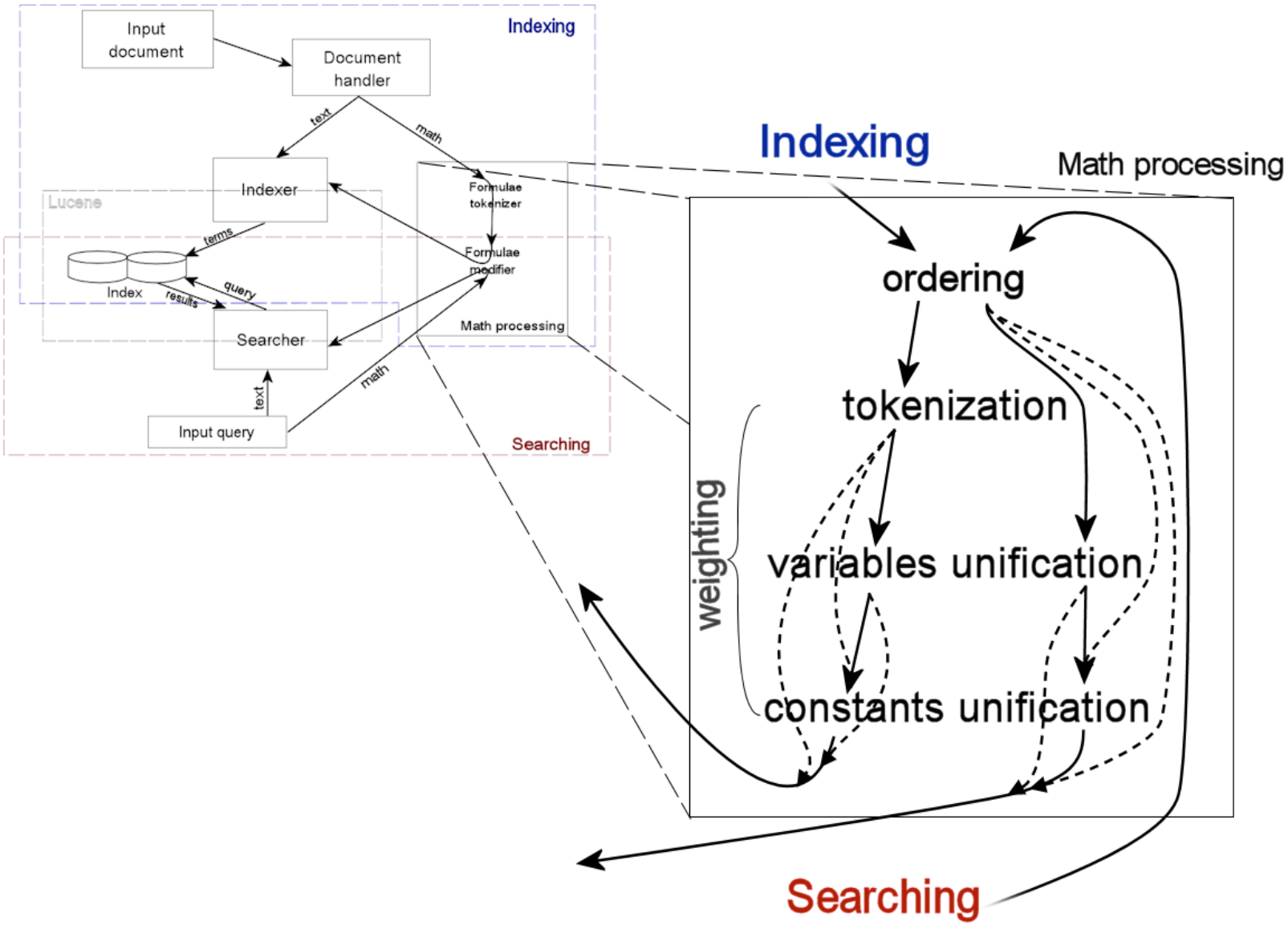
Math Indexer and Searcher MIaS — Features

- Inspired mostly by MathDex and EgoMath.
- Presentation and now also Content MathML.
- Allows *similarity* (not only exact match) between query and matched term, *distributional representation* of formulae.
  - Commutativity.
  - Unification of variables and constants.
  - Subformulae matching.
- Level of similarity calculation for expressions.
- Mixed mathematical-textual queries.
- Based on full text state of the art Apache Lucene core.

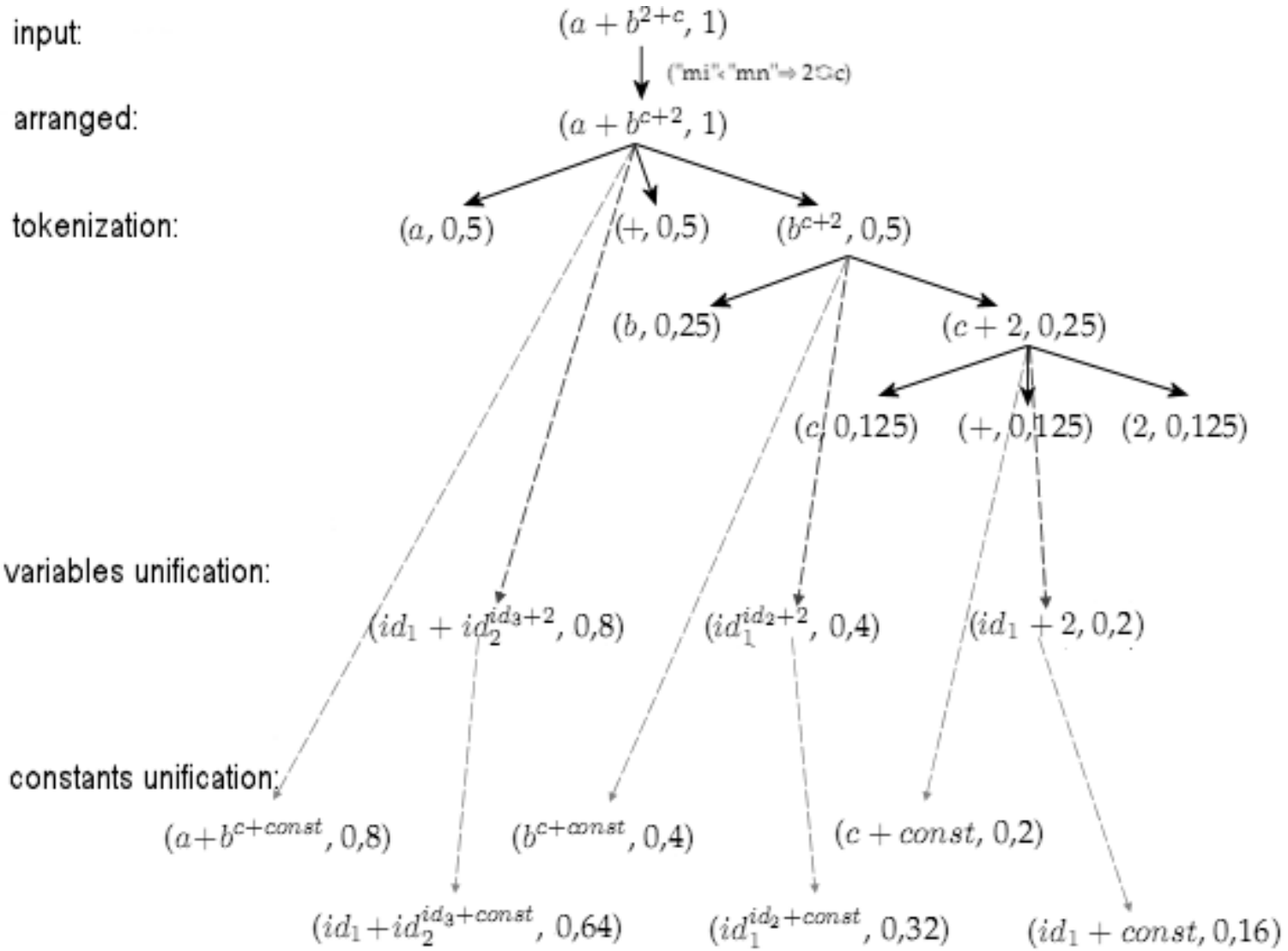
Math Indexer and Searcher — Overall Design



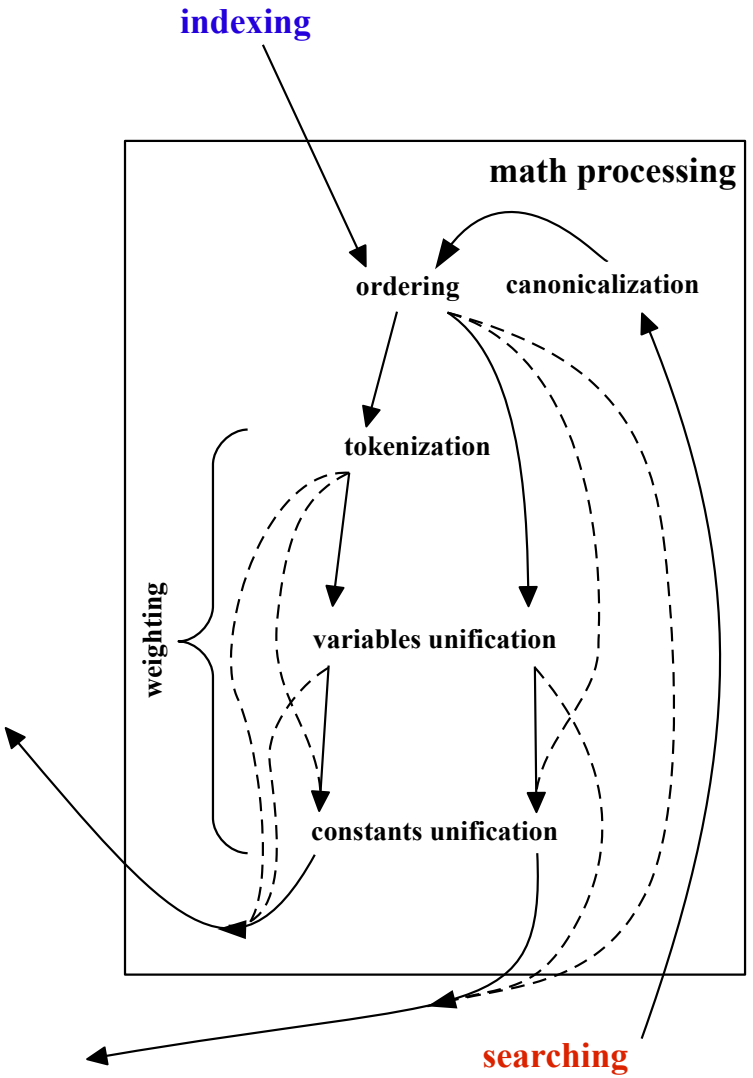
Math Indexer and Searcher — Math Workflow Design



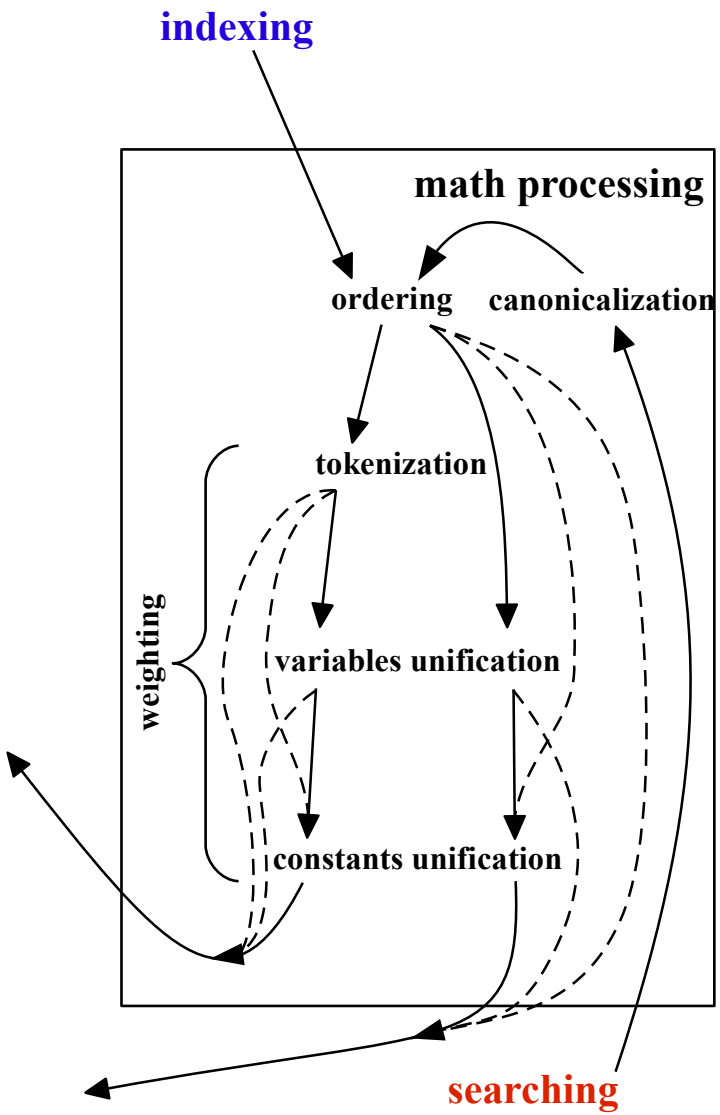
Formula Processing Weighting Example

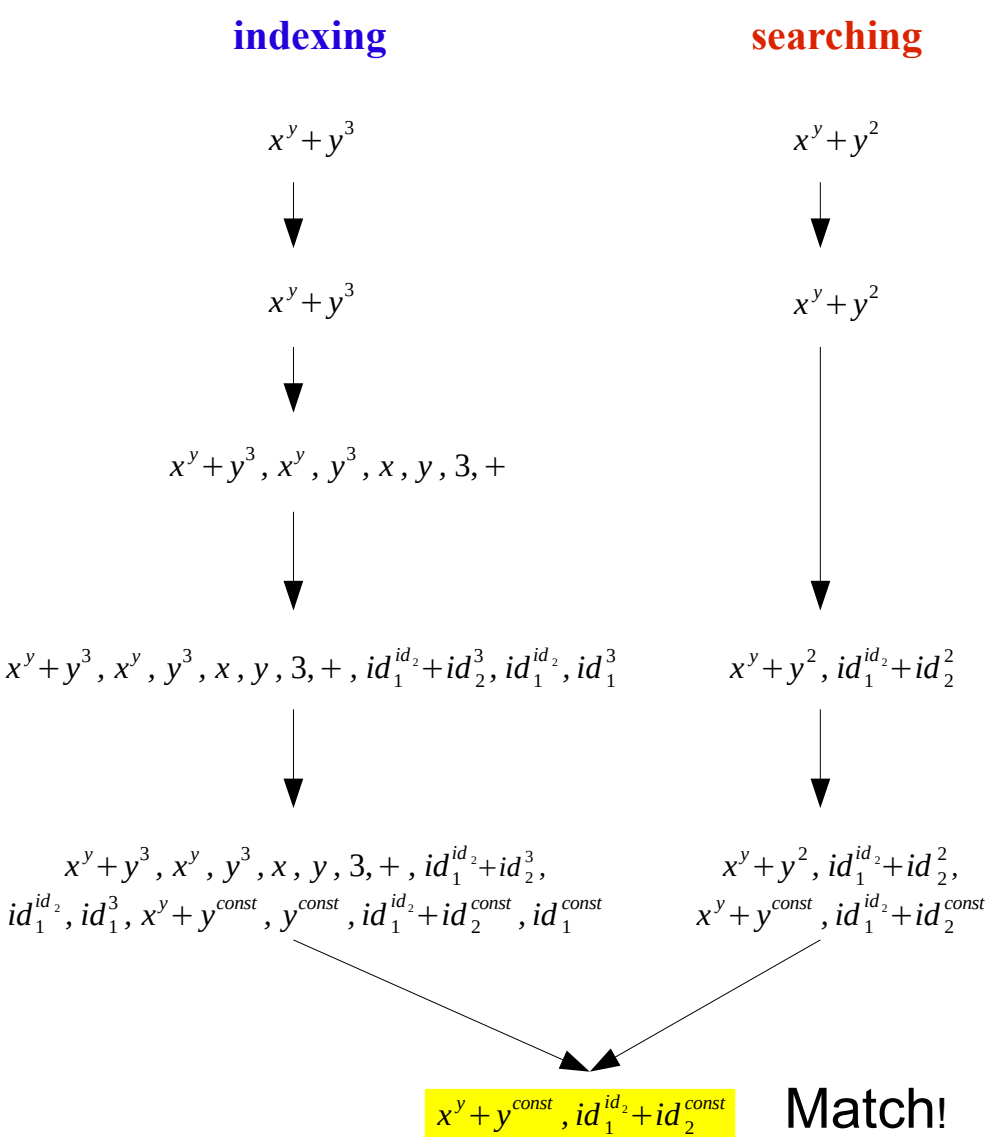


Math Formulae Indexing Processing



Example


indexing
math processing
ordering
canonicalization
tokenization
weighting
variables unification
constants unification
searching


indexing
searching
x^y+y^3
x^y+y^3
$x^y+y^3, x^y, y^3, x, y, 3, +$
$x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3, id_1^{id_2}, id_1^3$
$x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3, id_1^{id_2}, id_1^3, x^y+y^{const}, y^{const}, id_1^{id_2}+id_2^{const}, id_1^{const}$
x^y+y^2
x^y+y^2
$x^y+y^2, id_1^{id_2}+id_2^2$
$x^y+y^2, id_1^{id_2}+id_2^2, x^y+y^{const}, id_1^{id_2}+id_2^{const}$
$x^y+y^{const}, id_1^{id_2}+id_2^{const}$
Match!

Implementation

- Java
- Solr + Lucene.
- scalable: indexing $10^{10}+$ formulae without problems.
- Mathematical part implements Lucene's interface Tokenizer — able to integrate to any Solr/Lucene based system as DSpace, Elasticsearch…

Formulae Search Demonstration Comments

Demo web interface: https://mir.fi.muni.cz/webmias-ntcir/

- MathML/TEX input (LaTeXML for conversion to MathML).
- Canonicalization of the query – our own MathCanEval canonicalizer (developed by students as part of Dean's program at FI MU).
- Matched document snippet generation.
- MathJax for nicer math rendering and better portability.
- Snuggle TeX for on-the-fly as-you-type rendering.

All up and ready on the EuDML system: <http://eudml.org/search/>

MIaS4NTCIR: data indexing statistics

Table: Index statistics

Indexing times [min]		Index
Wall Clock	**CPU**	**size [GiB]**
1,940.0	3,413.55	68

Table: Formulae count statistics

	Formulae	
Documents	**Original**	**Indexed**
8,301,545	59,647,566	3,021,865,236

MIaS4NTCIR: Canonicalization

We have designed, implemented and continually improve a converter<https://mir.fi.muni.cz/mathml-normalization/> for *both* Presentation and Content MathML for this task.

MathCanEval application developed by Michal Růžička (lead), David Formánek, Dominik Szalai, Robert Šiška, Jakub Adler is designed and developed for evaluation of the canonicalizer.

MIaS4NTCIR: Canonicalization II

MathMLCanEval | Jmeno | Heslo | Prihlasit

Vzorec

Detaily vzorce

ID	7519
Uzivatel	admin
Zdrojovy dokument	ntcir-10
Konverzni program	LaTeXML
Poznamky	

MathML kod

```
<?xml version="1.0"?>
<math xmlns="http://www.w3.org/1998/Math/MathML" xmlns:dct="
  <semantics id="idp3882624">
    <mfrac id="idp3882752">
      <mrow id="idp3882880">
        <msup id="idp3883008">
          <mi id="idp3883136">e</mi>
          <mo id="idp3883392">+</mo>
        </msup>
        <mo id="idp3883648">+</mo>
        <msup id="idp3883904">
          <mi id="idp3884032">e</mi>
          <mo id="idp3884288">-</mo>
        </msup>
      </mrow>
      <mn id="idp3884544">2</mn>
    </mfrac>
    <annotation-xml id="idp3884800" encoding="MathML-Content"
      <apply id="idp3885168">
        <divide id="idp3885296"/>
        <apply id="idp3885424">
          <plus id="idp3885552"/>
          <apply id="idp3885680">
            <csymbol id="idp3885808" cd="ambiguous">superscr
            <ci id="idp3886368">e</ci>
            <plus id="idp3886624"/>
          </apply>
          <apply id="idp3886752">
            <csymbol id="idp3886880" cd="ambiguous">superscr
            <ci id="idp3887440">e</ci>
            <minus id="idp3887696"/>
          </apply>
        </apply>
        <cn id="idp3887824" type="integer">2</cn>
      </apply>
    </annotation-xml>
    <annotation id="idp3888352" encoding="application/x-tex"
  </semantics>
</math>
```

$$\frac{e^{+}+e^{-}}{2}$$

Provedene kanonikalizace

Cas	ID	Hash	Konfigurace	Doba behu
2014-06-02T19:27:24.358+02:00	6568	7065b77e57410ff12fd2911015b7c178fea6bdeb	1	2

Ekvivaletni formule

MathMLCanEval | Jmeno | Heslo | Prihlasit

Kanonikalizovany vzorec

Najit podobne

Podrobnosti

ID	6568
Puvodni vzorce	7519
Doba behu	2

Anotace

MathML kod

Kanonizovany kod | Puvodni kod | Rozdil

```
<math display="inline" alttext="\frac{e^{+}+e^{-}}{2}" id="idp3881904"
xmlns:acm="http://alpha.planetmath.org/onto/acm"
xmlns:msc="http://alpha.planetmath.org/onto/msc"
xmlns:arx="http://arxmliv.kwarc.info/ns/" xmlns:dct="http://purl.org/dc/terms/"
xmlns="http://www.w3.org/1998/Math/MathML">
 <semantics id="idp3882624">
  <mfrac id="idp3882752">
   <mrow id="idp3882880">
    <msup id="idp3883008">
     <mi id="idp3883136">e</mi>
     <mo id="idp3883392">+</mo>
    </msup>
    <mo id="idp3883648">+</mo>
    <msup id="idp3883904">
     <mi id="idp3884032">e</mi>
     <mo id="idp3884288">-</mo>
    </msup>
   </mrow>
   <mn id="idp3884544">2</mn>
  </mfrac>
  <annotation-xml encoding="MathML-Content" id="idp3884800">
   <apply id="idp3885168">
    <divide id="idp3885296"></divide>
    <apply id="idp3885424">
     <plus id="idp3885552"></plus>
     <apply id="idp3885680">
      <csymbol cd="ambiguous" id="idp3885808">superscript</csymbol>
      <ci id="idp3886368">e</ci>
      <plus id="idp3886624"></plus>
     </apply>
     <apply id="idp3886752">
      <csymbol cd="ambiguous" id="idp3886880">superscript</csymbol>
      <ci id="idp3887440">e</ci>
      <minus id="idp3887696"></minus>
     </apply>
    </apply>
    <cn type="integer" id="idp3887824">2</cn>
   </apply>
  </annotation-xml>
  <annotation encoding="application/x-tex" id="idp3888352">\frac{e^{+}+e^{-}}
{2}</annotation>
 </semantics>
</math>
```

```
<math xmlns:msc="http://alpha.planetmath.org/onto/msc"
xmlns:dct="http://purl.org/dc/terms/"
xmlns:arx="http://arxmliv.kwarc.info/ns/"
xmlns:acm="http://alpha.planetmath.org/onto/acm"
xmlns="http://www.w3.org/1998/Math/MathML">
 <semantics>
  <mfrac>
   <mrow>
    <msup>
     <mi>e</mi>
     <mo>+</mo>
    </msup>
    <mo>+</mo>
    <msup>
     <mi>e</mi>
     <mo>-</mo>
    </msup>
   </mrow>
   <mn>2</mn>
  </mfrac>
  <annotation-xml encoding="MathML-Content">
   <apply>
    <divide></divide>
    <apply>
     <plus></plus>
     <apply>
      <csymbol>superscript</csymbol>
      <ci>e</ci>
      <plus></plus>
     </apply>
     <apply>
      <csymbol>superscript</csymbol>
      <ci>e</ci>
      <minus></minus>
     </apply>
    </apply>
    <cn type="integer">2</cn>
   </apply>
  </annotation-xml>
  <annotation encoding="application/x-tex">\frac{e^{+}+e^{-}}
{2}</annotation>
 </semantics>
</math>
```

diff view generated by jsdifflib

MIaS4NTCIR: Representation of Math, Structures, (Meaning) for Indexing

Concepts of *similarity* and *distributional representations* are central in the design of MIaS. Every formulae is represented in the index as a *set of weighted tokens (subformulae, features)* that grab both structure and content of indexed mathematical formulae. The weighting is computed via small set of rules reflecting similarity distance of indexed tokens to the original formulae: the more similar is token to the original (in size, variable naming, constants used, …), the higher weighting score is stored in the index for a token. On average, currently the formulae representation is distributed over about 30 indexed weighted tokens.

MIaS4NTCIR: Query Expansion

subquery 1 (the original query):	f_1	f_2	k_1	k_2	k_3
subquery 2:	f_1	f_2	k_1	k_2	
subquery 3:	f_1	f_2	k_1		
subquery 4:	f_1	f_2			
subquery 5:	f_1		k_1	k_2	k_3
subquery 6:			k_1	k_2	k_3

Figure: Complete sequence of subqueries derived from the original user's query

Results merging, finally.

Query Expansion Results' Insight

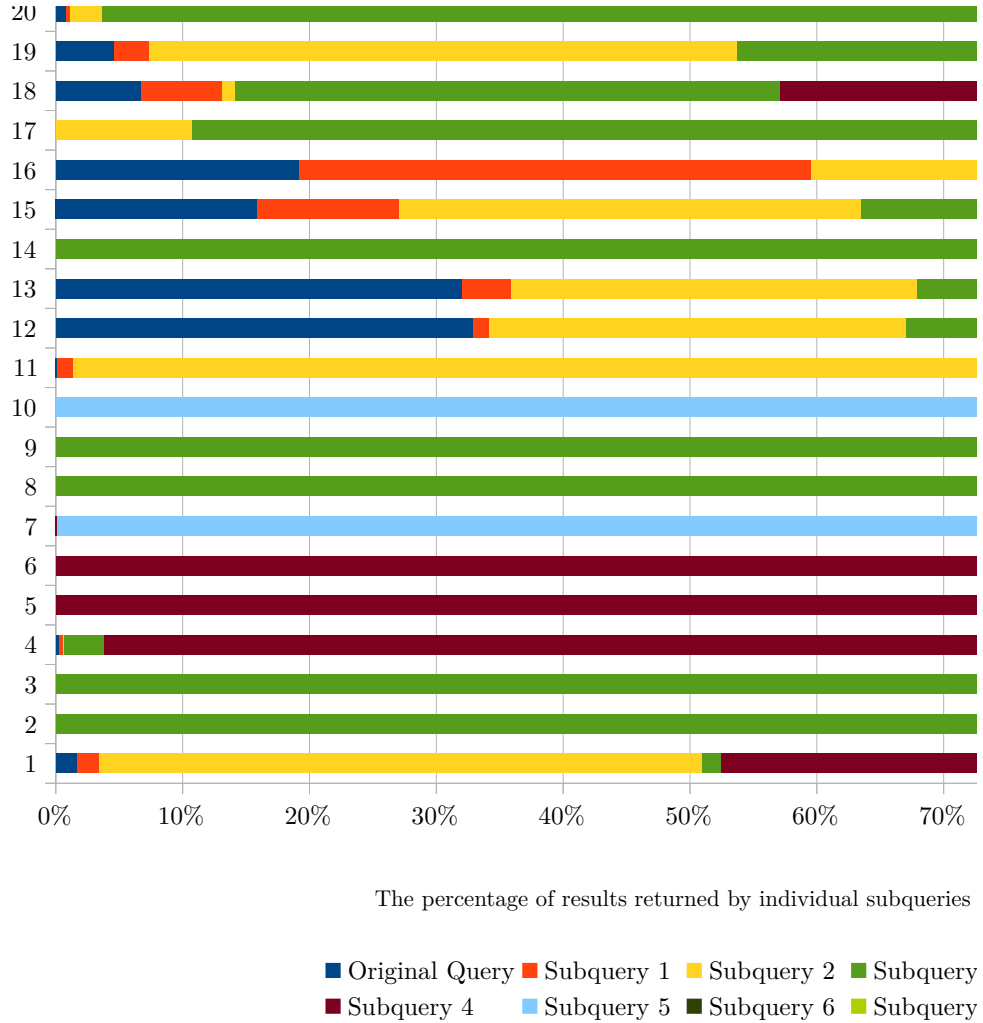


Figure: Relative number of results found using different subqueries for every query in CMath run

MIaS Results: 4 runs PMath, CMath, PCMath, TEX

Table: Results of submitted runs with Relevance Level ≥ 3 (Relevant). Main task team rank is in [ ] for our best runs (in bold).

	PMath	**CMath**	**PCMath**	**TEX**
MAP avg	0.3073	**0.3630 [1]**	0.3594	0.3357
P@10 avg	0.3040	**0.3520 [1]**	0.3480	0.3380
P@5 avg	0.5120	**0.5680 [1]**	0.5560	0.5400

Table: Results of submitted runs with Relevance Level ≥ 1 (Partially Relevant). Number in [ ] is team rank of all runs.

	PMath	**CMath**	**PCMath**	**TEX**
MAP avg	0.2557	**0.2807 [2]**	0.2799	0.2747
P@10 avg	0.5020	0.5440	**0.5520 [1]**	0.5400
P@5 avg	0.8440	**0.8720 [2]**	0.8640	0.8480

Test Hardware Description

- Physical server (no virtualization).
- Shared with other research groups (with no resource reservations/prioritizations).
  - Jobs running with low priorities.
  - Unpredictable load on CPUs/memory/disks.
  - There almost always is significant load on the server.
- 8 × Intel Xeon X7560 @ 2.27 GHz (64 cores).
- 448 GiB of RAM.
- 8 × 300 GiB 10k RPM disk, organized in hardware RAID10.
- More info: <https://www.fi.muni.cz/tech/unix/aura.xhtml> (Sorry, in Czech only, please use Google Translator.)

Software Description

- Red Hat Enterprise Linux Server 6.
- Java 7.
- Lucene.

Reproducibility of the Setup

- Should be possible:
  - Open-source tools.
  - No commercial software in use.
  - Available data.
- Costs:
  - Hardware.
  - Power.
  - Human resources.

MIaS Wikipedia Task Performance

- Indexing time:
  - 26 min.
- Response times:
  - min: ~0 sec.
  - max: 3.489 sec.
  - avg: 0.177 sec.
- Overall time including input/output processing:
  - 4.33 hours.
  - However, *huge* (?85 %?) portion of the time is wasted by Perl libxml module processing and pretty printing of huge XML files that is useful to have nice logs but is not really necessary.

MIaS Wikipedia Task Indexing

- Indexing size:
  - ~750 MiB
- Index contents:
  - Processed formulae in the M-Terms format.
  - Analyzed text, full texts are *not* included.

MIaS Wikipedia Task Results

- Topics with results:
  - 75 (CMath run)
- Average position:
  - 65 correct results in top 1000
  - 64 correct results in top 100
  - 58 correct results in top 20
  - 56 correct results in top 10
  - 53 correct results in top 5
  - 52 correct results in top 4
  - 50 correct results in top 3
  - 48 correct results in top 2
  - 46 correct results in top 1

MIaS Wikipedia Task Content Topics

- According to Moritz. ‘MIRMU: The only team that has submitted an actual run. All the other teams seem to have done that manually.’
- Completely the same fully automatic system used for the main NTCIR Math Task and Wikipedia subtask.
  - Only different data.
  - No tuning or modifications for the Wikipedia task.
- Input Content MathML was transformed to the format of the main NTCIR math task.
  - Manually added Presentation MathML and TeX representation of the data.
  - Performed all the four runs (CMath, PMath, PCMath, TeX) similarly to the main task.

MIaS Wikipedia Task Content Topics Results

- No results for query 1: $\frac{1}{\langle x \rangle} \leq \langle \frac{1}{x} \rangle$
- Few results for query 2: $f_{xy} = f_{yx}$
  - 41 results for TeX run (the best one).
  - 0 results for CMath run (the worst one).

MIaS Wikipedia Task Content Topics – Noticeable Results

- No results for CMath run at all.
  - Better canonicalization of the hand made Content MathML probably needed.
- Query 2: $f_{xy} = f_{yx}$
  - Subformula search works (math.21697.28):
    - $\sigma_{rr} = \frac{1}{r}\frac{\partial\varphi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\varphi}{\partial\theta^2}$; $\sigma_{\theta\theta} = \frac{\partial^2\varphi}{\partial r^2}$; $\sigma_{r\theta} = \sigma_{\theta r} = -\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial\varphi}{\partial\theta}\right)$
  - Correct result (math.2085.76)?
    - $\sqrt{\rho_n}R_{n\to m}\frac{1}{\sqrt{\rho_m}} = H_{nm}$
  - Majority of the results are simple formulae like:
    - $G_{\mu\nu} = G_{\nu\mu}$

Semantic Gap between Lexical Surface of the Text and its Meaning in [M]IR

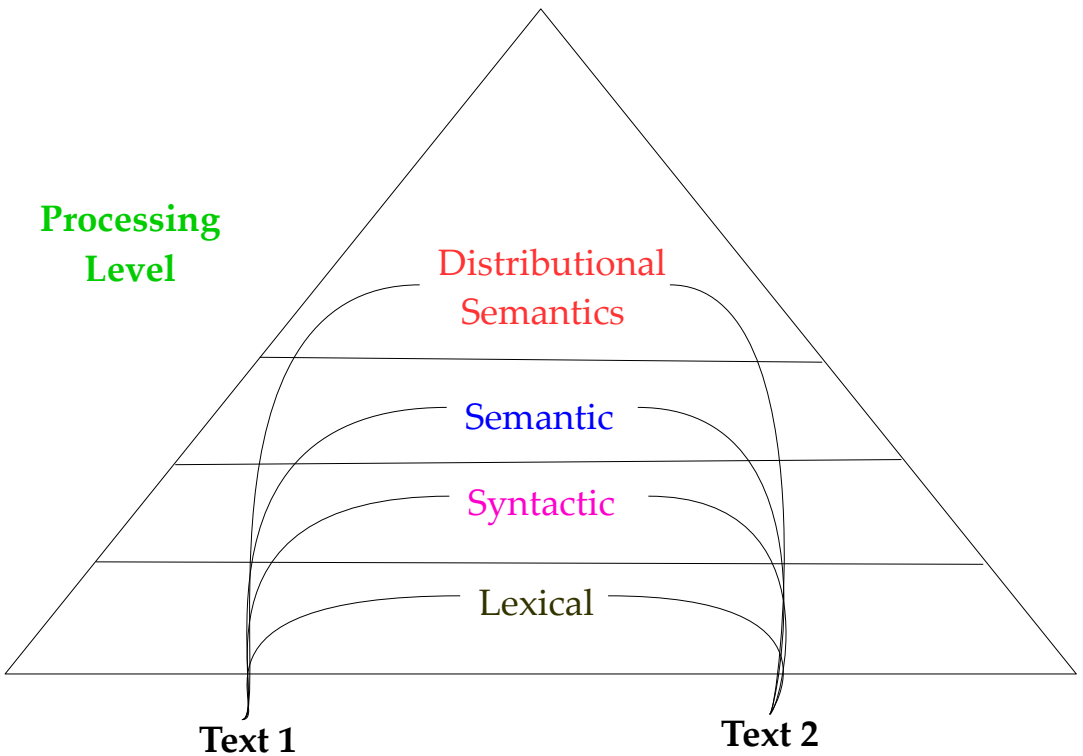
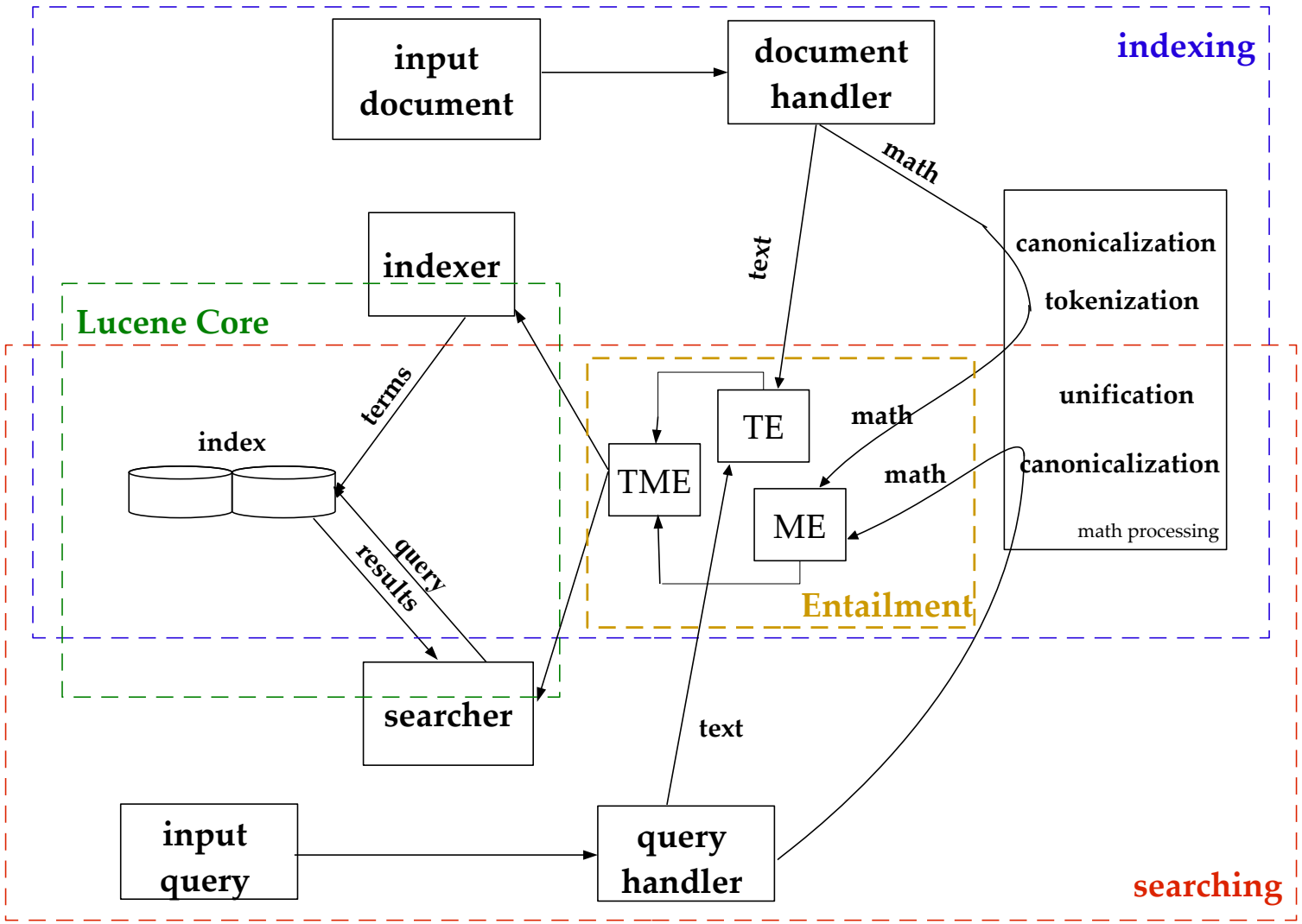
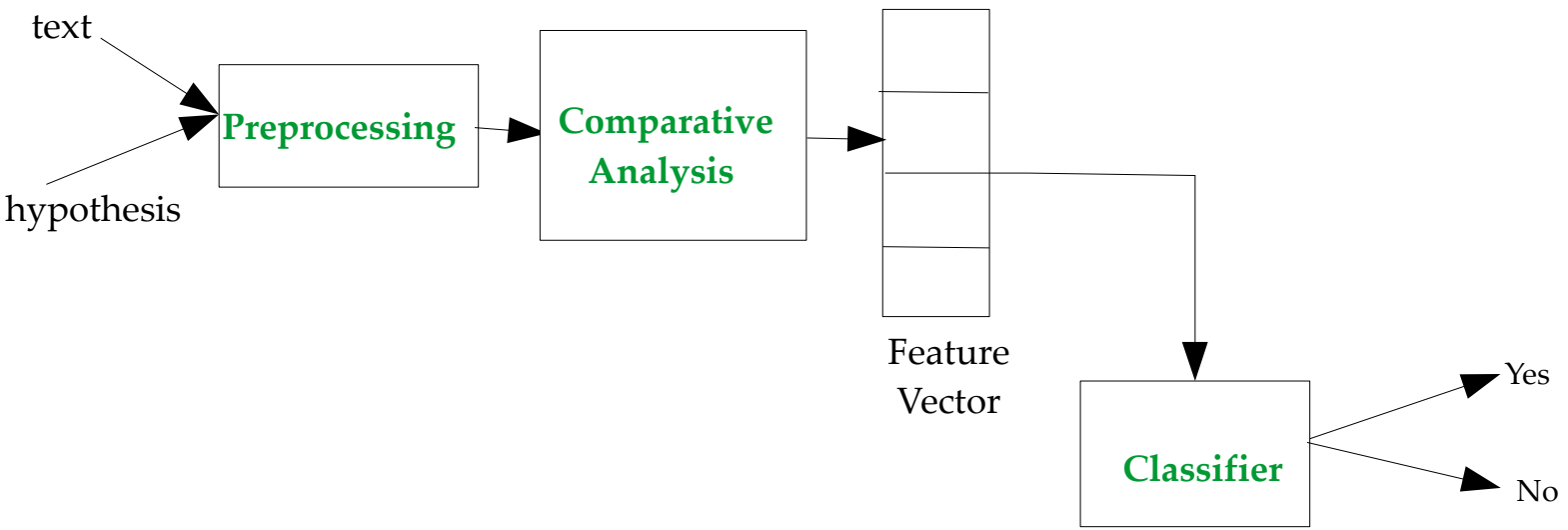


Figure: Natural language processing levels

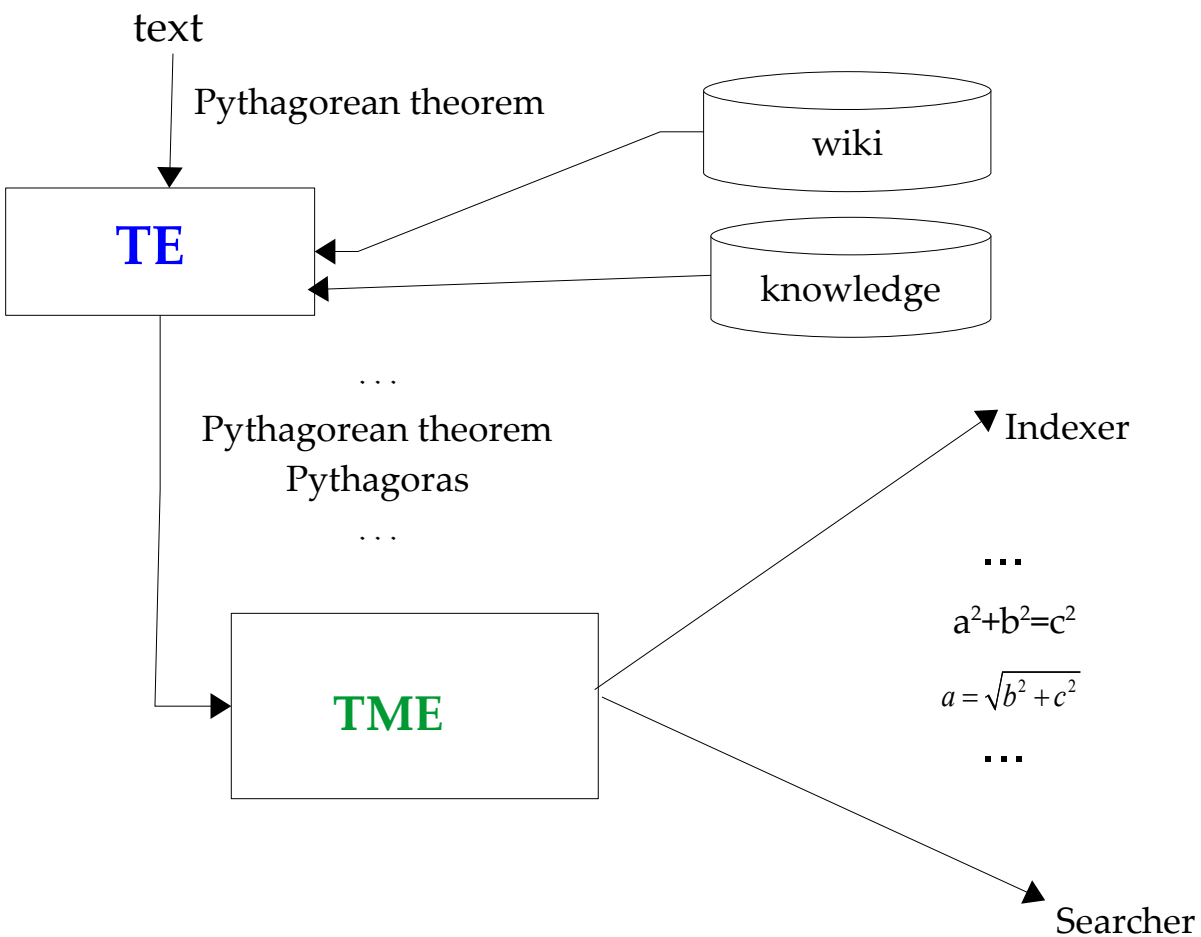
New MIaS Architecture with Textual and Math Entailment Modules



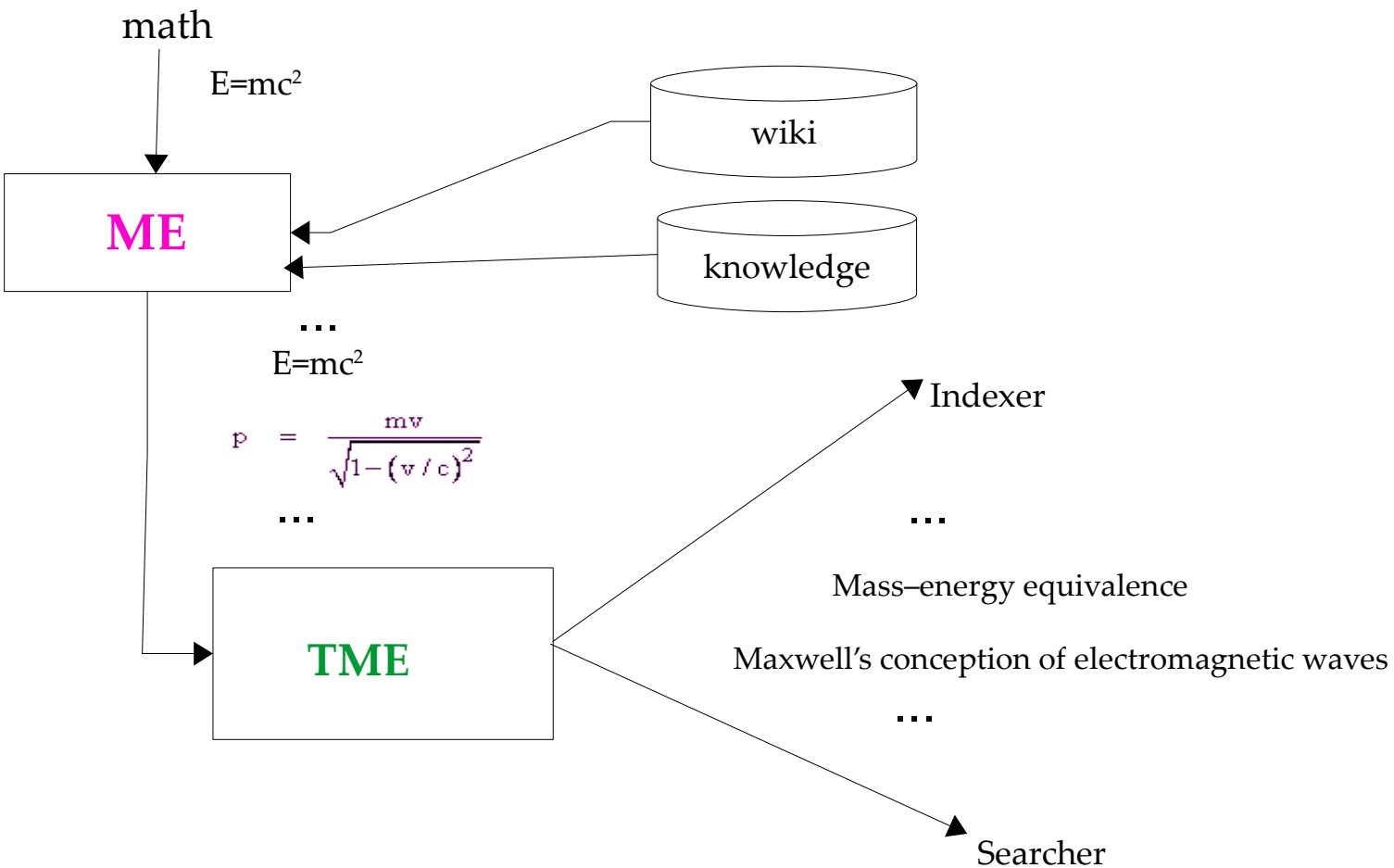
General Textual Entailment Architecture



Data Flow in TE and TME Modules



Data Flow in ME and TME Modules



Future work?

- Math entailment trained on Wikipedia math data.
- Full text mining in semantic direction (typesetting^{-1}), higher level NLP.
- Globalization (Google Scholar), deploying global knowledge bases.
- Personalization (up to the individual's preferences).
- Increase of automation and precision on semantic level.

Future Challenges

- Meaningful math-aware knowledge representation.
- Math entailment (Partha Pakray), ‘flexiformat’ processing, ‘canonicalization’ (?Strict CMathML) of math formulae.
- Math-aware corpora processing.
- Only then challenges as: multilingual math retrieval, MathML indexing and search, math common sense, text and math disambiguation and understanding, mathematical document classification, document similarity could be possible.

Challenge of Math-aware Distributional Semantics Processing

- Math-aware knowledge representation: handling abstractions, high-dimensional vector space representations?
- Math2vec? 'Smooth' vector space representation of math formulae learnt by recurrent neural network: `math2vec` aka `word2vec` (T. Mikolov from Brno, now Google), GloVe (Stanford's tool for distributional semantics), COMPOSES Semantic vectors (M. Baroni's way of distributional semantics).
- Hyper-lapsed vector space representation of documents (narrative qualities, rephrased plagiarism).

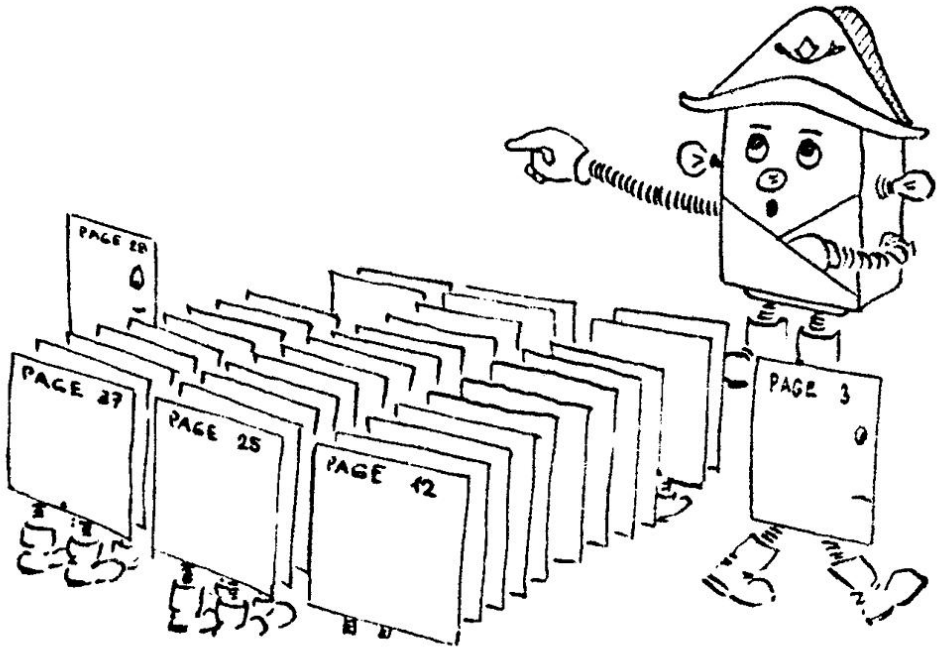
Challenge of Math-aware Corpora Processing and Tools

- Canonicalization of math formulae processing (MathCanEval).
- Switching between different levels of structured data.
- Tools adaptation (handling trees and abstractions), ideally on data acquired and tagged without supervision.

Challenge of Evaluation of Math Information Retrieval

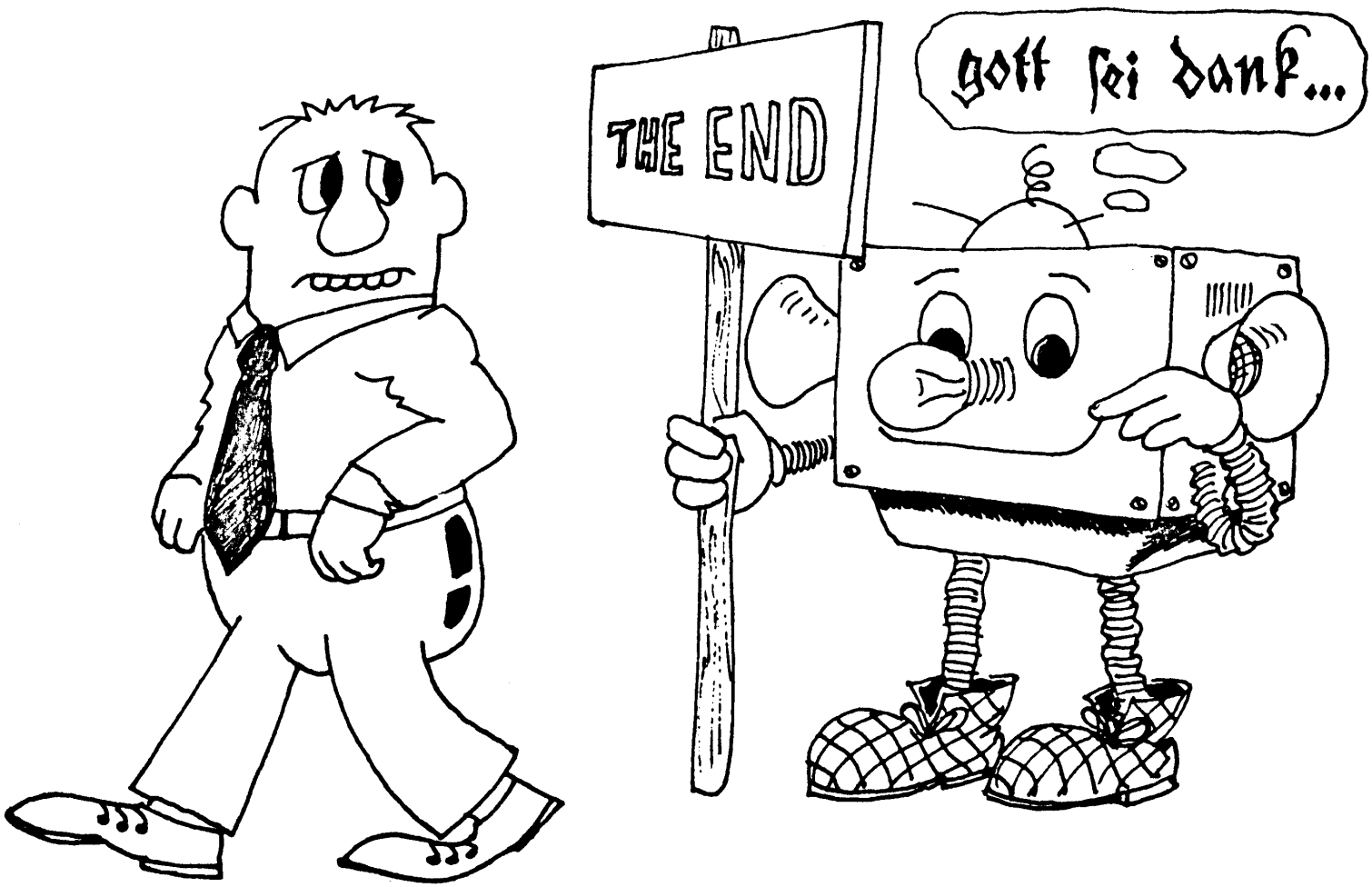
- What works in math-aware IR, UI, pragmatics.
- First MIR happening in 2012, now regular Math Tasks at NTCIR-10, NTCIR-11.
- Deploying MIaS and our tools in the future [?G]DML projects.

Acknowledgments and Questions?



Acknowledgements: EuDML and DML-CZ projects (funding), EuDML and DML-CZ colleagues, Martin Líška, Michal Růžička, Radim Řehůřek, David Formánek, Dominik Szalai, Robert Šiška, Jakub Adler, Partha Pakray, Radim Hatlapatka, Martin Jarmar, Maroš Kucbel, Zuzana Nevěřilová, Mirek Bartošek, Martin Šárfy, Vlastík Krejčíř, Petr Kovář, Vlastimil Dohnal, and many, many other authors and contributors of tools used.

That's it!




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>


Líška, Martin and Petr Sojka and Michal Růžička. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task. In Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Math Pilot Task. pp. 686-691. NII, Tokyo, 2013. PDF


D. Formánek, M. Líška, M. Růžička, and P. Sojka. Normalization of digital mathematics library content. In J. Davenport, J. Jeuring, C. Lange, and P. Libbrecht, editors, 24th OpenMath Workshop, 7th Workshop on Mathematical User Interfaces (MathUI), and Intelligent Computer Mathematics Work in Progress, number 921 in CEUR Workshop Proceedings, pp. 91–103, Aachen, 2012.


Sojka, Petr and Martin Líška. The Art of Mathematics Retrieval. In Matthew R. B. Hardy , Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View, CA, USA: ACM, 2011. p. 57–60. ISBN 978-1-4503-0863-2. <http://dx.doi.org/10.1145/2034691.2034703>


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>.


Credits for LDA pictures goes to David M. Blei.


Credits for illustrations goes to Jiří Franek.