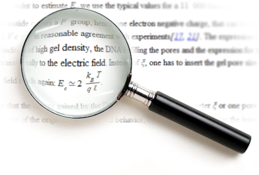
Maths Information Retrieval for Digital Libraries

Michal Růžička

Masaryk University, Faculty of Informatics, Brno, Czech Republic
mruzicka@mail.muni.cz

https://mir.fi.muni.cz/



FACULTAS ARTIS INFORMATICAE • UNIVERSITAS MASARYKIANA •

Illustrations by Jiří Franek.

Outline

1. Motivation
2. Math Indexer and Searcher
3. MathML Canonicalizer
4. My Goals
5. Evaluation
6. Future Works
7. Summary

Motivation

Motivation

- *Czech Digital Mathematics Library* (DML-CZ), `http://www.dml.cz/`
- *Centre de diffusion de revues académiques mathématiques* (CEDRAM), `http://www.numdam.org/`
- *Numérisation de documents anciens mathématiques* (NUMDAM), `http://www.cedram.org/`
- *Göttingen Göttinger Digitalisierungszentrum* (GDZ), `http://gdz.sub.uni-goettingen.de/`
- *Electronic Research Archive for Mathematics* (ERAM), `http://www.emis.de/projects/JFM/`
- *The Electronic Library of Mathematics* (ELibM), `http://siba-sinmemis.unile.it/ELibM.html`
- *Journal STORage* (JSTOR), `http://www.jstor.org/`
- *Project Euclid*, `http://projecteuclid.org/`
- *Russian Digital Mathematics Library* (RusDML), `http://www.rusdml.de/`
- *Polish Digital Mathematical Library* (DML-PL), `http://pldml.icm.edu.pl/`
- *Biblioteca Digital Española de Matemáticas* (DML-E), `http://dmle.cindoc.csic.es/`
- *Japanese Digital Mathematics Library* (DML-JP), `http://sparc1.math.sci.hokudai.ac.jp/dmljp/`
- *Riviste Elettroniche Italiane di Matematica* (REIM), `http://siba2.unile.it/sinm/reim/`
- *Biblioteca Digitale Italiana di Matematica* (bdim), `http://www.bdim.eu/`

Motivation (cont.)

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

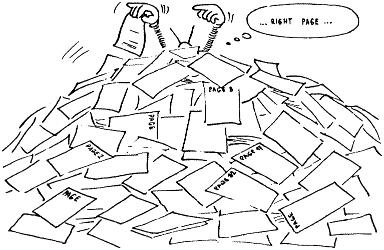
A: **'Math formulae search.'**

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

Topics/questions:
1) What functionality and incentives would made a working mathematician to login and use a modern DML as EuDML?
2) What are the most decisive factors on this crossroad?

Motivation (cont.)

- DML without maths-aware search support is an oxymoron.

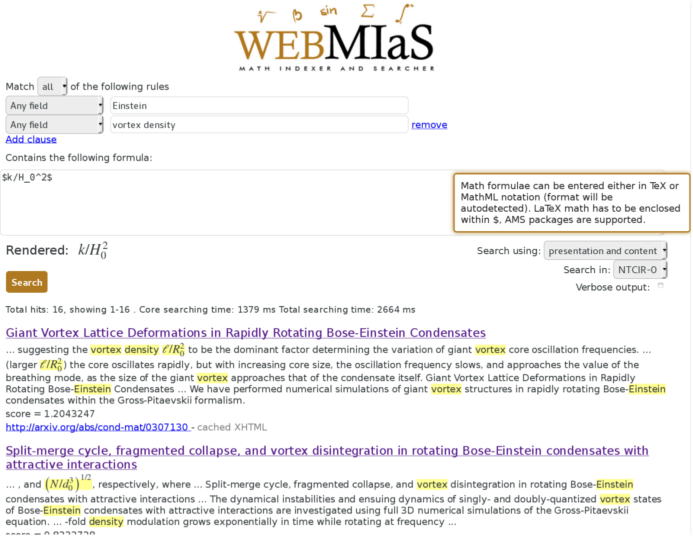


- Simple search based on text keywords is not appropriate or sufficient for mathematical contents.

Outline

1. Motivation
2. **Math Indexer and Searcher**
3. MathML Canonicalizer
4. My Goals
5. Evaluation
6. Future Works
7. Summary

Math Indexer and Searcher (MIaS)



Match all of the following rules

Any field: Einstein

Any field: vortex density remove

Contains the following formula:

```
$k/H_0^2$
```

Math formulae can be entered either in TeX or MathML notation (format will be autodetected). LaTeX math has to be enclosed within $, AMS packages are supported.

Rendered: k/H_0^2

Search using: presentation and content

Search in: NTCIR-0

Verbose output: □

Total hits: 16, showing 1-16 . Core searching time: 1379 ms Total searching time: 2664 ms

Giant Vortex Lattice Deformations in Rapidly Rotating Bose-Einstein Condensates

… suggesting the vortex density ℓ/R_0^2 to be the dominant factor determining the variation of giant vortex core oscillation frequencies. … (larger ℓ/R_0^2) the core oscillates rapidly, but with increasing core size, the oscillation frequency slows, and approaches the value of the breathing mode, as the size of the giant vortex approaches that of the condensate itself. Giant Vortex Lattice Deformations in Rapidly Rotating Bose-Einstein Condensates … We have performed numerical simulations of giant vortex structures in rapidly rotating Bose-Einstein condensates within the Gross-Pitaevskii formalism.

score = 1.2043247

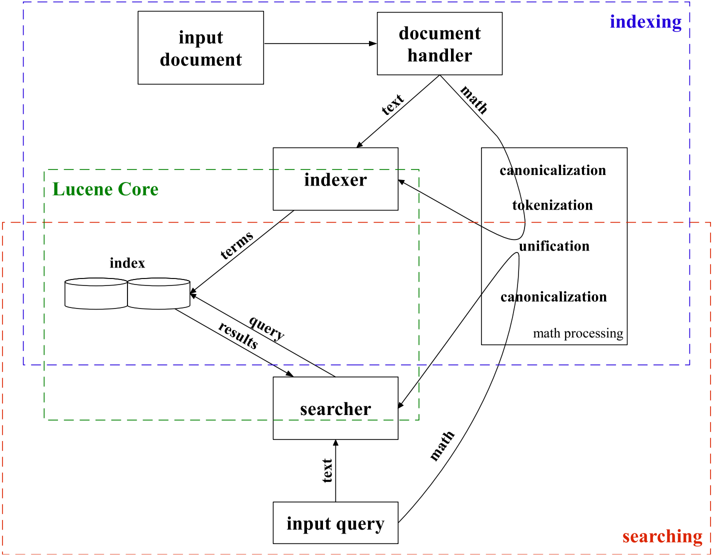
http://arxiv.org/abs/cond-mat/0307130 - cached XHTML

Split-merge cycle, fragmented collapse, and vortex disintegration in rotating Bose-Einstein condensates with attractive interactions

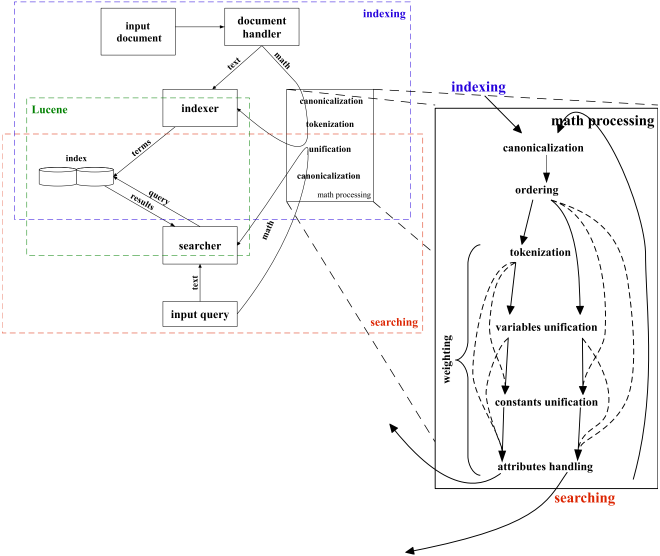
… , and $\left(N/d_0^3\right)^{1/2}$, respectively, where … Split-merge cycle, fragmented collapse, and vortex disintegration in rotating Bose-Einstein condensates with attractive interactions … The dynamical instabilities and ensuing dynamics of singly- and doubly-quantized vortex states of Bose-Einstein condensates with attractive interactions are investigated using full 3D numerical simulations of the Gross-Pitaevskii equation. … -fold density modulation grows exponentially in time while rotating at frequency …

score = 0.8232728

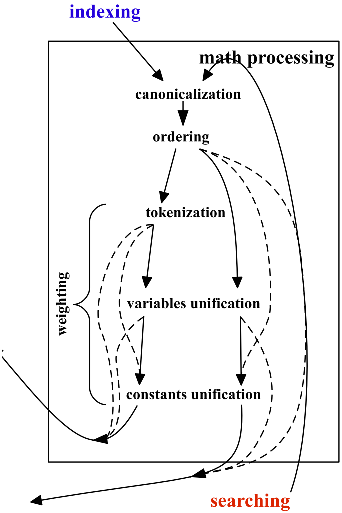
MIaS Design


indexing
input document
document handler
text
math
indexer
Lucene Core
canonicalization
tokenization
unification
canonicalization
math processing
terms
index
query
results
searcher
text
math
input query
searching

MIaS Design


indexing
input document
document handler
text
math
indexer
Lucene
canonicalization
tokenization
unification
canonicalization
math processing
terms
index
query
results
searcher
math
text
input query
searching
indexing
math processing
canonicalization
ordering
tokenization
variables unification
weighting
constants unification
attributes handling
searching

MIaS Indexing and Searching Example



indexing

math processing

canonicalization

ordering

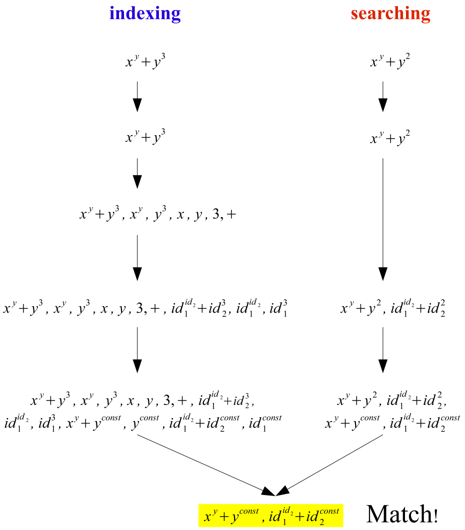
tokenization

weighting

variables unification

constants unification

searching



indexing

x^y+y^3

x^y+y^3

$x^y+y^3,\ x^y,\ y^3,\ x,\ y,\ 3,\ +$

$x^y+y^3,\ x^y,\ y^3,\ x,\ y,\ 3,\ +,\ id_1^{id_2}+id_2^3,\ id_1^{id_2},\ id_1^3$

$x^y+y^3,\ x^y,\ y^3,\ x,\ y,\ 3,\ +,\ id_1^{id_2}+id_2^3,\ id_1^{id_2},\ id_1^3,\ x^y+y^{const},\ y^{const},\ id_1^{id_2}+id_2^{const},\ id_1^{const}$

searching

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!

Outline

1. Motivation
2. Math Indexer and Searcher
3. **MathML Canonicalizer**
4. My Goals
5. Evaluation
6. Future Works
7. Summary

Sources of MathML in Digital Libraries

- ‘Hand made’
- Tralics
- LaTeXML
- InftyReader
- MaxTract
- MATLAB
- Wolfram Alpha
- ...

‘Hand made’ $x^2 + y^2$ MathML

```
<math xmlns=’http://www.w3.org/1998/Math/MathML’>
    <msup>
        <mi>x</mi><mn>2</mn>
    </msup>
    <mo>+</mo>
    <msup>
        <mi>y</mi><mn>2</mn>
    </msup>
</math>
```

Sources of MathML in Digital Libraries

- ‘Hand made’
- Tralics
- LaTeXML
- InftyReader
- MaxTract
- MATLAB
- Wolfram Alpha
- …

Matlab $x^2 + y^2$ MathML

```
generate::MathML(x^2 + y^2,
                 Content = FALSE, Annotation = FALSE)
<math xmlns=’http://www.w3.org/1998/Math/MathML’>
  <mrow xref=’No7’>
    <msup xref=’No3’>
      <mi xref=’No1’>x</mi>
      <mn xref=’No2’>2</mn>
    </msup>
    <mo>+</mo>
    <msup xref=’No6’>
      <mi xref=’No4’>y</mi>
      <mn xref=’No5’>2</mn>
    </msup>
  </mrow>
</math>
```

Sources of MathML in Digital Libraries

- ‘Hand made’
- Tralics
- LaTeXML
- InftyReader
- MaxTract
- MATLAB
- Wolfram Alpha
- ...

LaTeXML $x^2 + y^2$ MathML

```
<math xmlns=”http://www.w3.org/1998/Math/MathML”
   alttext=”x^{2}+y^{2}” display=”inline”>
   <semantics>
      <mrow>
         <msup><mi>x</mi><mn>2</mn></msup>
         <mo>+</mo>
         <msup><mi>y</mi><mn>2</mn></msup>
      </mrow>
      <annotation encoding=”application/x-tex”>
         x^{2}+y^{2}
      </annotation>
   </semantics>
</math>
```

Sources of MathML in Digital Libraries

- ‘Hand made’
- Tralics
- LaTeXML
- InftyReader
- MaxTract
- MATLAB
- Wolfram Alpha
- …

InftyReader $x^2 + y^2$ MathML

```
<math xmlns=”http://www.w3.org/1998/Math/MathML”>
   <msup>
      <mi mathvariant=”italic”>x</mi>
      <mrow>
         <mn mathvariant=”normal”>2</mn>
      </mrow>
   </msup>
   <mo mathvariant=”normal”>+</mo>
   <msup>
      <mi mathvariant=”italic”>y</mi>
      <mrow>
         <mn mathvariant=”normal”>2</mn>
      </mrow>
   </msup>
</math>
```

MathML Canonicalizer

- Our own MathML canonicalization tool.
- The main design imperatives:
  - Modularity,
  - simplicity,
  - extensibility, and
  - flexibility.
- The speed of the canonicalization application is also a critical parameter.
  - In our MREC corpora there is 168,000,000 formulae to canonicalize.

MathML Canonicalizer Use Cases

<mphantom> Omission

```
<mfrac>
  <mrow>
    <mi> x </mi>
    <mo> + </mo>
    <mi> y </mi>
    <mo> + </mo>
    <mi> z </mi>
  </mrow>
  <mrow>
    <mi> x </mi>
    <mphantom>
      <mo> + </mo>
      <mi> y </mi>
    </mphantom>
    <mo> + </mo>
    <mi> z </mi>
  </mrow>
</mfrac>
```

```
<mfrac>
  <mrow>
    <mi> x </mi>
    <mo> + </mo>
    <mi> y </mi>
    <mo> + </mo>
    <mi> z </mi>
  </mrow>
  <mrow>
    <mi> x </mi>




    <mo> + </mo>
    <mi> z </mi>
  </mrow>
</mfrac>
```

MathML Canonicalizer Use Cases

Unnecessary Attributes

```
<mfrac linethickness=”2”
       bevelled=”true”>
  <mi> a </mi>
  <mi> b </mi>
</mfrac>
```

```
<mfrac>
  <mi> a </mi>
  <mi> b </mi>
</mfrac>
```

MathML Canonicalizer Use Cases

<mrow> Minimizing

```
<msqrt>
  <mrow>
    <mo> - </mo>
    <mn> 1 </mn>
  </mrow>
</msqrt>
```

```
<msqrt>

  <mo> - </mo>
  <mn> 1 </mn>

</msqrt>
```

MathML Canonicalizer Use Cases

Unifying Fences

```
<mfenced open=”[”>

  <mi> x </mi>

  <mi> y </mi>

</mfenced>
```

```
<mrow>
  <mo> [ </mo>
  <mrow>
    <mi> x </mi>
    <mo> , </mo>
    <mi> y </mi>
  <mrow>
  <mo> ) </mo>
</mrow>
```

MathML Canonicalizer Use Cases

Sub-/Superscripts Handling

```
<msubsup>
  <mi> x </mi>

  <mn> 1 </mn>

  <mn> 2 </mn>
</msubsup>
```

```
<msup>
  <msub>
    <mi> x </mi>
    <mn> 1 </mn>
  </msub>
  <mn> 2 </mn>
</msup>
```

MathML Canonicalizer Use Cases

Applying Functions

```
<mi> f </mi>
<mo> &#x2061; </mo>
<mrow>
  <mo> ( </mo>
  <mi> x </mi>
  <mo> ) </mo>
</mrow>
```

```
<mi> f </mi>
<mrow>
  <mo> ( </mo>
  <mi> x </mi>
  <mo> ) </mo>
</mrow>
```

MathML Canonicalizer Use Cases

Applying Functions

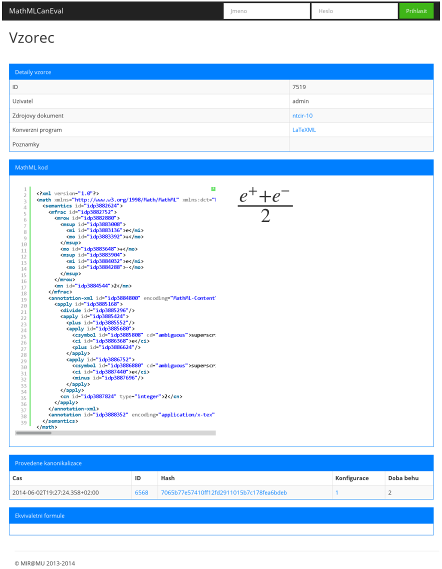
```
<mi> sin </mi>

<mo> &#x2061; </mo>

<mi> x </mi>
```

```
<mi>sin</mi>
<mrow>
  <mo>(</mo>
  <mi>x</mi>
  <mo>)</mo>
</mrow>
```

MathML Canonicalizer Web Evaluation Application

- JUnit testing seems not to be enough.
- Collaboration-enabled evaluation system needed:
    - Visualization of test data collection.
    - History of canonicalization results on the same data with different version of the Canonicalizer.
    - Annotations on the results.
    - Statistics.
    - Coverage of all the mark-up of the MathML standard.

MathML Canonicalizer Web Evaluation Application



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>
```

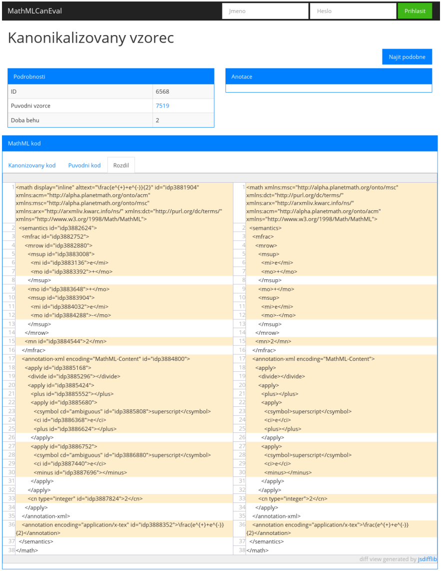
$$\frac{e^{+}+e^{-}}{2}$$

Provedene kanonikalizace

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

Ekvivaletni formule

© MIR@MU 2013-2014



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>
```

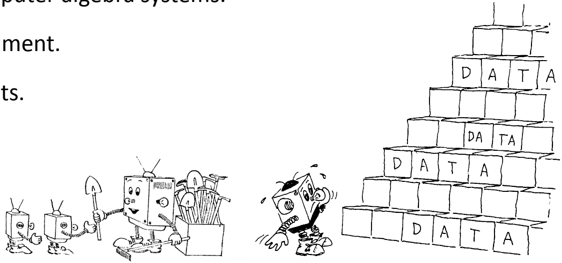
diff view generated by jsdifflib

Outline

1. Motivation
2. Math Indexer and Searcher
3. MathML Canonicalizer
4. My Goals
5. Evaluation
6. Future Works
7. Summary

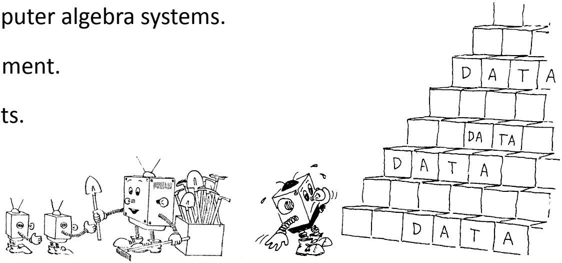
Objectives

- Improvements of the relevance of the results of our math-aware search engine.
    - MathML Normalization.
    - Classification of identifiers.
    - Context driven search.
    - Involvement of computer algebra systems.
    - Image search experiment.
    - Ranking of the results.



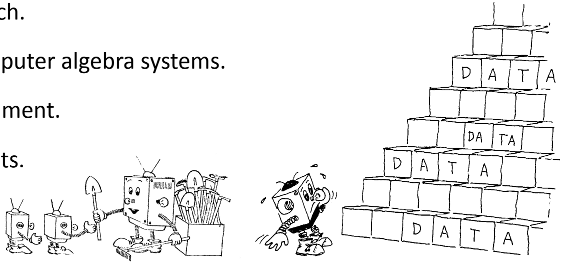
Objectives

- Improvements of the relevance of the results of our math-aware search engine.
    - MathML Normalization.
        - Canonicalization of both Presentation and Content MathML.
    - Classification of identifiers.
    - Context driven search.
    - Involvement of computer algebra systems.
    - Image search experiment.
    - Ranking of the results.



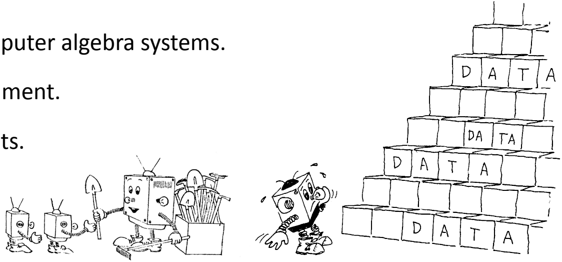
Objectives

- Improvements of the relevance of the results of our math-aware search engine.
    - MathML Normalization.
    - Classification of identifiers.
        - To mark particular identifiers as variable name, function name, and so on.
        - Start from the metadata available for the documents.
    - Context driven search.
    - Involvement of computer algebra systems.
    - Image search experiment.
    - Ranking of the results.



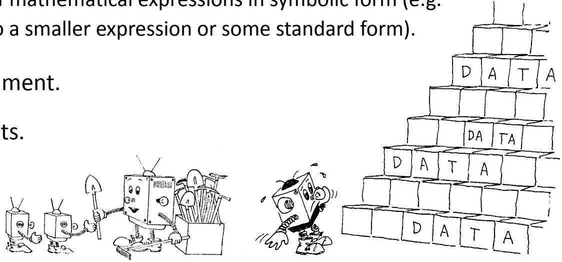
Objectives

- Improvements of the relevance of the results of our math-aware search engine.
    - MathML Normalization.
    - Classification of identifiers.
    - Context driven search.
        - Exploitation of the metadata available for the documents.
        - Exploitation of the users' inputs.
    - Involvement of computer algebra systems.
    - Image search experiment.
    - Ranking of the results.



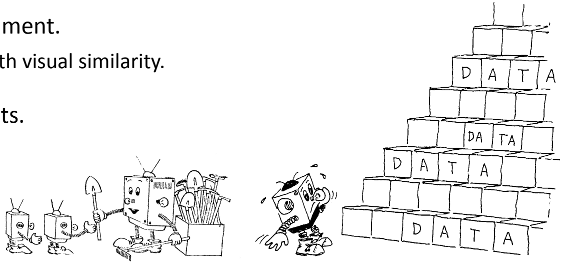
Objectives

- Improvements of the relevance of the results of our math-aware search engine.
    - MathML Normalization.
    - Classification of identifiers.
    - Context driven search.
    - Involvement of computer algebra systems.
        - Manipulation of mathematical expressions in symbolic form (e.g. simplification to a smaller expression or some standard form).
    - Image search experiment.
    - Ranking of the results.



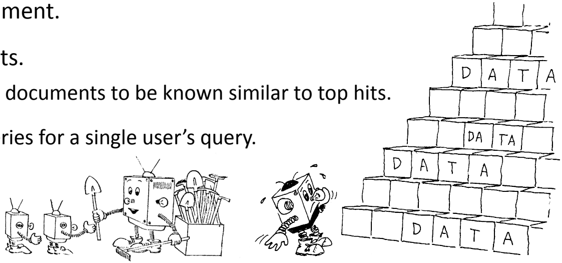
Objectives

- Improvements of the relevance of the results of our math-aware search engine.
  - MathML Normalization.
  - Classification of identifiers.
  - Context driven search.
  - Involvement of computer algebra systems.
  - Image search experiment.
    - Experiments with visual similarity.
  - Ranking of the results.



Objectives

- Improvements of the relevance of the results of our math-aware search engine.
  - MathML Normalization.
  - Classification of identifiers.
  - Context driven search.
  - Involvement of computer algebra systems.
  - Image search experiment.
  - Ranking of the results.
    - Score bonus for documents to be known similar to top hits.
    - Internal subqueries for a single user's query.



Outline

1. Motivation
2. Math Indexer and Searcher
3. MathML Canonicalizer
4. My Goals
5. Evaluation
6. Future Works
7. Summary

NTCIR-10 Math Task

- The fist (pilot) year of the math task event last year (i.e. 2013).
- Formula search and Full-text search.
  - 4 runs submitted – differ in query language.
    - PMath – Run #1.
    - CMath – Run #2.
    - PCMath – Run #3.
    - $\TeX$ – Run #4.
- Open Information Retrieval.
  - 1 run submitted – $\TeX$ + text mixed queries.

NTCIR-10 Math Task Results

Table 1: Result metrics for submitted runs in Formula Search with Relevance Level ≥ 3 (Relevant)

Metric	**Run 1**	**Run 2**	**Run 4**
P-10 avg	0.105	0.191	**0.219**
P-5 avg	0.133	0.229	**0.276**
MAP avg	0.060	0.112	**0.127**
Precision	0.109 (64/589)	**0.185** (92/496)	0.123 (96/778)

Table 2: Result metrics for submitted runs in Formula Search with Relevance Level ≥ 1 (Partially Relevant)

Metric	**Run 1**	**Run 2**	**Run 4**
P-10 avg	0.143	0.214	**0.267**
P-5 avg	0.181	0.267	**0.343**
MAP avg	0.066	0.081	**0.100**
Precision	0.148 (87/589)	**0.232** (115/496)	0.161 (125/778)

NTCIR-11 Math Task

- A greater number of participants.
    - Increase from 6 to 8.
- Only one type of queries.
    - 50 queries, each
        - 1–4 formulae,
        - 1–4 keyphrases.
- Our results submitted for judgement in June 2014.

NTCIR-11 Math Task: Our Investigation of Our Results

Improper query conversion

Index: `\operatorname{Im}P^{+}_{\Gamma}=C_{\mu}^{+}(\Gamma)`

Query: `ImP^{+}_{\gamma}=C^{+}_{\mu}(\gamma)`

Index

```
 ...
<mrow>
  <mo>Im</mo>
  <mo></mo>
  <msup>
 ...
```

Query

```
 ...
<mrow>
  <mi>I</mi>
  <mi>m</mi>
  <msup>
 ...
```

NTCIR-11 Math Task: Our Investigation of Our Results

Substructure difference tolerance should be improved

```
...
  <mrow>
    QUERY-FORMULA-SUBPART-1
    [[ INDEX
        <mrow>
          <mo>∫</mo>
          <mrow>
    || QUERY
        <mi>o</mi>
    ]]
    QUERY-FORMULA-SUBPART-2
    INDEX-FORMULA-SUPPLEMENT
        </mrow>
      </mrow>
    </mrow>
  </mrow>
...
```

NTCIR-11 Math Task: Our Investigation of Our Results

\qvar{} handling

Original task query:

```
\qvar{S}=-\qvar{T}_{\qvar{p}}\int\qvar{d}^{\qvar{p}+1}
  \qvar{x}\sqrt{\qvar{g}}
```

Index: `S=-T_{p}\int d^{p+1}x\sqrt{-g}`

Query: `S=-T_{p}\int d^{p+1}x\sqrt{g}`

`\sqrt{g}` does not match `\sqrt{-g}`.

NTCIR-11 Math Task: Our Investigation of Our Results

Unification would be helpful

Query:

```
\qvar{x}\frac{\qvar{y}}{\qvar{z}}-\qvar{u}\frac{\qvar{v}}{\qvar{w}}
```

$$x\frac{y}{z} - u\frac{v}{w}$$

Matches:

$$\begin{aligned}\{q_s, q_r\} &= \int dx \int dy \, \{A(x,\mu)^p, A(y,\nu)^q\} \; \Big|_{\mu^{s+1}\nu^{r+1}} \\ &= pq\,\mu\nu \int dx\, A(\mu)^p (A(\nu)^q)' \left[\left\{\frac{s}{p}\nu - \frac{r}{q}\mu\right\}\frac{1}{\mu - \nu} + \frac{1}{h}\frac{rs}{pq}\right] \; \Big|_{\mu^{s+1}\nu^{r+1}}\end{aligned}$$

Does not match:

$$\zeta \sim c_1\frac{\delta\rho_\sigma}{\rho_\sigma} - c_2\frac{\delta H_{\mathrm{osc}}}{H_{\mathrm{osc}}}$$

NTCIR-11 Math Task: Our Investigation of Our Results

- Combination of both formulae and text keywords in one query is important.
- Multiple subqueries derived from the original query with result lists merging turned out to be very useful.
  - One-by-one removal of the keywords and formulae.
  - “Strip-merging” of the results of the subqueries.
  - Further investigation of the best strategies needed.
    - Subqueries with subformulae?
    - Different strategy for merging of results?

Outline

1. Motivation
2. Math Indexer and Searcher
3. MathML Canonicalizer
4. My Goals
5. Evaluation
6. Future Works
7. Summary

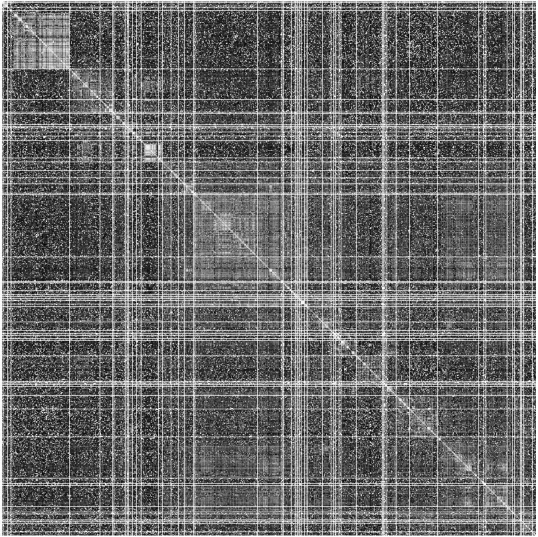
Future Works

- Long-term goals:
  - All the objectives mentioned earlier. ☺
  - Whole arXiv indexing for MIaS search.
  - Evaluation.
- Short-term goals:
  - Improved strategies for MIaS internal subqueries derived from a single user's query.
  - Integration of MIaS with DML-CZ DSpace.
  - Exploitation of Gensim-Math-computed document similarities to improve of ranking results in MIaS.
  - Evaluation, Evaluation, Evaluation.

Future Works – Gensim Math Document Similarities

- Gensim by Radim Řehůřek is “the most robust, efficient and hassle-free piece of software to realize unsupervised semantic modelling from plain text”: `http://radimrehurek.com/gensim/`
- We are experimenting with correlation of document similarities based on Mathematics Subject Classification (MSC) vs. document similarities based on formulae and other prominent parts (title, authors, abstract…) of the documents.
- Visualized similarity matrices.

Future Works – Gensim Math Document Similarities



Method: TfIdf-LSI; Weighted MTerms: true; MTerm Weight Conversion: 1

Future Works – Gensim Math Document Similarities

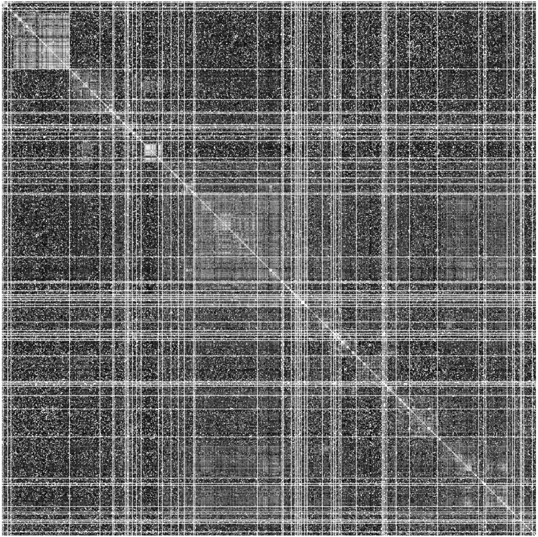
- MSC sorted documents in columns/rows, white lines separate MSC codes with different two characters (top category).
- All the documents compared each other.
  - Grayscale level indicates simarity of the document on the row to the document in the column.

white	Very similar.
black	Very different.

  - Different weighting (or even omitting) of tokens from various metadata fields.
  - Vector space model transformation method:

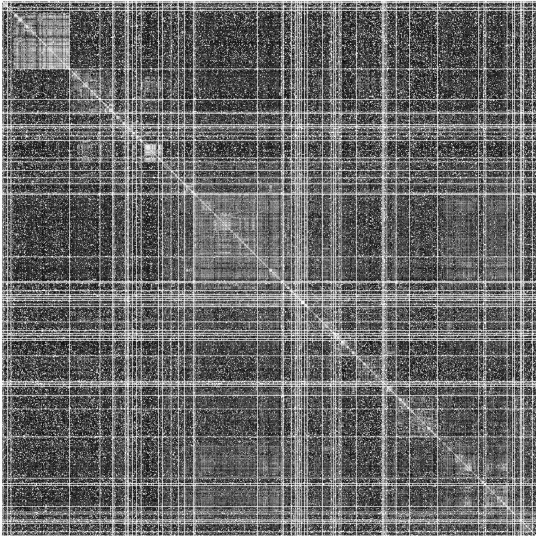
TfIdF-LSI	Inverse Document Frequency wrapped by Latent Semantic Indexing
LDA	Latent Dirichlet Allocation
...	...

Future Works – Gensim Math Document Similarities



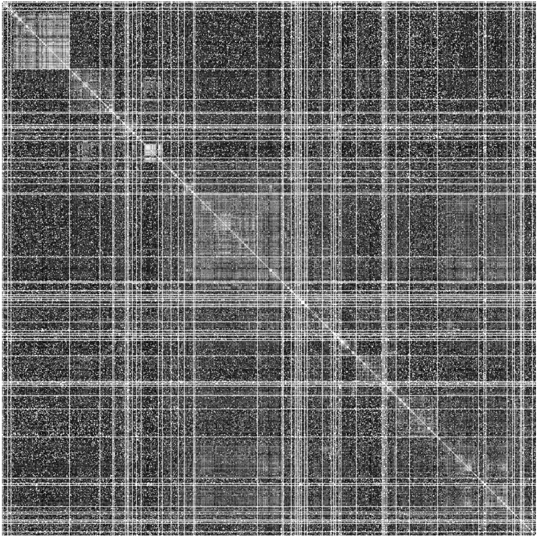
Method: TfIdf-LSI; Weighted MTerms: true; MTerm Weight Conversion: 1

Future Works – Gensim Math Document Similarities



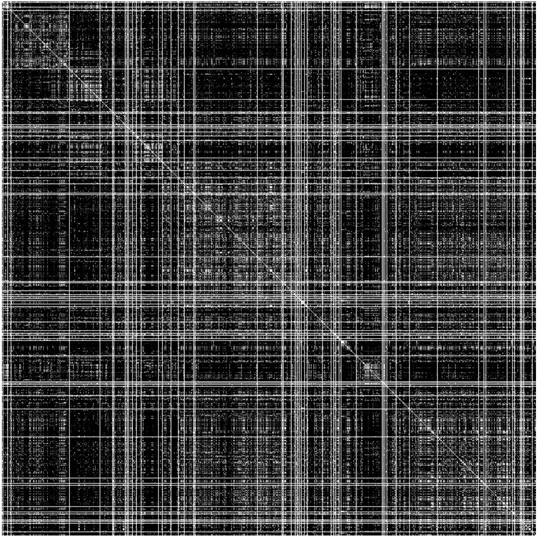
Method: TfIdf-LSI; Weighted MTerms: true; MTerm Weight Conversion: trunc(3.9 * mtermWeight)

Future Works – Gensim Math Document Similarities



Method: TfIdf-LSI; Weighted MTerms: false

Future Works – Gensim Math Document Similarities



Method: LDA; Weighted MTerms: false

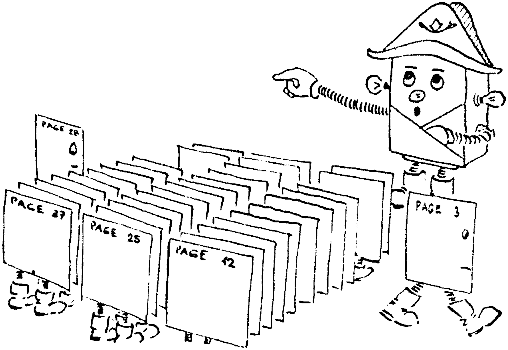
Future Works – Gensim Math Document Similarities

- Gensim by Radim Řehůřek is “the most robust, efficient and hassle-free piece of software to realize unsupervised semantic modelling from plain text”: `http://radimrehurek.com/gensim/`
- We are experimenting with correlation of document similarities based on Mathematics Subject Classification (MSC) vs. document similarities based on formulae and other prominent parts (title, authors, abstract…) of the documents.
- Visualized similarity matrices.
  - How to compute similarity of these matricies rigorously?
    - Canonical-correlation analysis (CCA)?

Summary

- We have our own math-aware search engine.
- We have a lot of possible improvements in our minds.
- We are interested in evaluation a lot.

Questions?



Illustrations by Jiří Franek.

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. doi:10.1145/2034691.2034703.

LÍŠKA, Martin, Petr SOJKA and Michal RŮŽIČKA. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task. In Noriko Kando, Kazuaki Kishida. Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan, 2013. s. 686-691, 6 s. ISBN 978-4-86049-062-1.

LÍŠKA, Martin, Petr SOJKA, Michal RŮŽIČKA and Peter MRAVEC. Web Interface and Collection for Mathematical Retrieval : WebMIaS and MREC. In Petr Sojka, Thierry Bouche. DML 2011: Towards a Digital Mathematics Library. Brno: Masaryk University, 2011. p. 77–84. ISBN 978-80-210-5542-1.

FORMÁNEK, David, Martin LÍŠKA, Michal RŮŽIČKA and Petr SOJKA. Normalization of Digital Mathematics Library Content. CEUR Workshop Proceedings, Aachen, 2012, vol. 921, October, p. 91–103. ISSN 1613-0073.

LÍŠKA, Martin, Petr SOJKA, Michal RŮŽIČKA and Peter MRAVEC. Web Interface and Collection for Mathematical Retrieval : WebMIaS and MREC. In Petr Sojka, Thierry Bouche. DML 2011: Towards a Digital Mathematics Library. Brno: Masaryk University, 2011. p. 77–84. ISBN 978-80-210-5542-1.

ŘEHŮŘEK, Radim and Petr SOJKA. Software Framework for Topic Modelling with Large Corpora. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010. p. 46–50. ISBN 2-9517408-6-7.