Search over corpora on mathematical texts

Peter Mravec
<256444@mail.muni.cz>

6th October 2010
Motivation

- Existing mathematical database.
  - ArXiv
  - Mathdex
- Their problem.
  - Full text search
- Continue in Martin Liška’s Bachelor’s thesis
Motivation

- Existing mathematical database.
  - ArXiv
  - Mathdex

- Their problem.
  - Full text search

- Continue in Martin Liška’s Bachelor’s thesis
Motivation

- Existing mathematical database.
  - ArXiv
  - Mathdex
- Their problem.
  - Full text search
- Continue in Martin Liška’s Bachelor’s thesis
About corpus

• Build corpus
  • MatCo - Mathematical Corpus
  • ArXiv.org
    • Physics
    • Mathematics
    • Computer Science
    • Quantitative Biology
    • Quantitative Finance
    • Statistics
About corpus

- Build corpus
- MatCo - Mathematical Corpus
  - ArXiv.org
    - Physics
    - Mathematics
    - Computer Science
    - Quantitative Biology
    - Quantitative Finance
    - Statistics
About corpus

• Build corpus

• MatCo - Mathematical Corpus

• ArXiv.org
  • Physics
  • Mathematics
  • Computer Science
  • Quantitative Biology
  • Quantitative Finance
  • Statistics
About corpus

- Build corpus
- MatCo - Mathematical Corpus
- ArXiv.org
  - Physics
  - Mathematics
  - Computer Science
  - Quantitative Biology
  - Quantitative Finance
  - Statistics
Download files

Figure: Download process.
ArXMLiv

- Transformation $\TeX$ documents to XML documents with MathML.
- Almost 500 000 documents.
ArXMLiv

- Transformation $\LaTeX$ documents to XML documents with MathML.
- Almost 500 000 documents.
### ArXMLiv

<table>
<thead>
<tr>
<th>return value</th>
<th>count</th>
<th>result</th>
<th>count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>unknown</td>
<td>697</td>
<td>none</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>no_latex</td>
<td>8986</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>missing_errlog</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fatal_error</td>
<td>30156</td>
<td>incomplete</td>
<td>35932</td>
<td>6.82</td>
</tr>
<tr>
<td>timeout</td>
<td>5776</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>error</td>
<td>37770</td>
<td>complete with errors</td>
<td>138755</td>
<td>26.32</td>
</tr>
<tr>
<td>missing_macros</td>
<td>100985</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>warning</td>
<td>286946</td>
<td>success</td>
<td>352408</td>
<td>66.86</td>
</tr>
<tr>
<td>no_problems</td>
<td>65462</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search over corpora on mathematical texts

6th October 2010
Corpus purpose

- Create index with Lucene and Manatee & Bonito.
- Compare:
  - hits
  - speed
  - performance
  - efectivity
  - advantage/disadvantage
Corpus purpose

- Create index with Lucene and Manatee & Bonito.
- Compare:
  - hits
  - speed
  - performance
  - efectivity
  - advantage/disadvantage
Corpus purpose

- Create index with Lucene and Manatee & Bonito.
- Compare:
  - hits
  - speed
  - performance
  - efectivity
  - advantage/disadvantage
Summary

• Build corpus of mathematical texts.
• Compare indexing methods.
• Suggest a solution for searching in EuDML-CZ repository.
Summary

- Build corpus of mathematical texts.
- Compare indexing methods.
- Suggest a solution for searching in EuDML-CZ repository.
Summary

• Build corpus of mathematical texts.
• Compare indexing methods.
• Suggest a solution for searching in EuDML-CZ repository.
Questions?

Search over corpora on mathematical texts

6th October 2010
References

Mgr. Vítězslav Dostál:
*Indexace matematických textů v digitální matematické knihovně.*
Brno, 2009.

Bc. Martin Líška
*Vyhledávání v matematickém textu*
Brno, 2010

Bc. Petr Kišš
*Portabilní fulltext pro CD/DVD*
Brno, 2006

Bc. Marek Chrenko
*Portabilní fulltext pro CD/DVD*
Brno, 2009

Cornell University Library
*ArXiv.org*
<http://arxiv.org/>

Knowledge Adaptation and Reasoning for Content
*The arXMLiv Developer Portal*
<https://trac.kwarc.info/arXMLiv/>