Search over corpora on mathematical texts

Peter Mravec

<256444@mail.muni.cz>

6th October 2010

EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

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

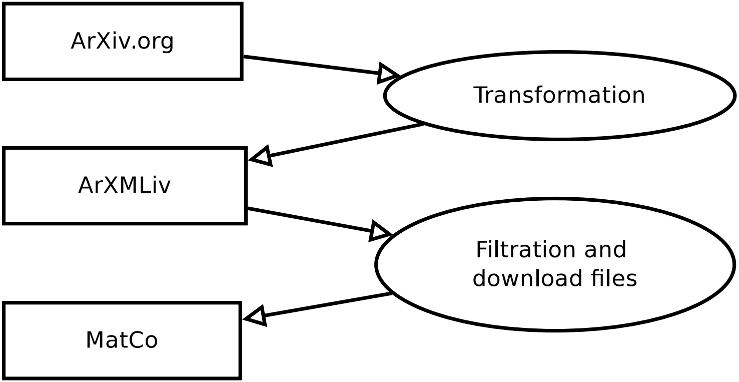


Figure: Download proces.

ArXMLiv

- Transformation TEX documents to XML documents with MathML.
- Almost 500 000 documents.

ArXMLiv

- Transformation TEX documents to XML documents with MathML.
- Almost 500 000 documents.

ArXMLiv

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

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.

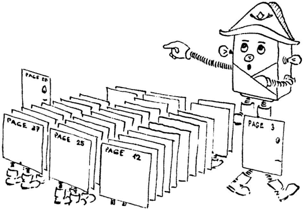
Summary

- 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?



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/>