Mathematical Indexing and Querying

Martin Líška

<255768@mail.muni.cz>

28th April 2010

EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Contents

- Motivation
- Mathematical notations
- Existing mathematical search engines
- Development of a mathematical indexing and searching engine
- Conclusion

Motivation

- Usefulness and importance of search engines
- Conventional search engines not applicable in the environment of a digital mathematics library
- Different rules for mathematics
  - different notation for equivalent formulae
  - same notation for different formulae

Motivation

- Usefulness and importance of search engines
- Conventional search engines not applicable in the environment of a digital mathematics library
- Different rules for mathematics
  - different notation for equivalent formulae
  - same notation for different formulae

Motivation

- Usefulness and importance of search engines
- Conventional search engines not applicable in the environment of a digital mathematics library
- Different rules for mathematics
  - different notation for equivalent formulae
  - same notation for different formulae

Mathematical notations

- TEX/LATEX
  - High quality typesetting
  - Simple, straightforward mathematics markup
  - Popular among scientist and in academic environment
  - Describes only syntax
- MathML
  - XML based, WWW standard developed by W3C consortium
  - Machine to machine math communication
  - Presentation MathML – syntaxt
  - Content MathML – semantics

Mathematical notations

- TEX/LATEX
  - High quality typesetting
  - Simple, straightforward mathematics markup
  - Popular among scientist and in academic environment
  - Describes only syntax
- MathML
  - XML based, WWW standard developed by W3C consortium
  - Machine to machine math communication
  - Presentation MathML – syntaxt
  - Content MathML – semantics

Mathematical notation II

- OpenMath
  - Emerging standard for describing semantics of mathematical objects
  - Strong relationship with MathML
  - Content dictionaries – for storing semantics
    - Basic CD is compatible with mathematical concepts of Content MathML
- OMDoc
  - Semantics-oriented mathematical representation format
  - Covers the whole range of written mathematics
  - Three levels:
    - Object level – formulae in Content MathML or OpenMath
    - Statement level – definitions, theorems, proofs, examples and the relations between them
    - Theory level – set of contextually related statements, OMDoc theories are compatible to OpenMath content dictionaries

Mathematical notation II

- OpenMath
  - Emerging standard for describing semantics of mathematical objects
  - Strong relationship with MathML
  - Content dictionaries – for storing semantics
    - Basic CD is compatible with mathematical concepts of Content MathML
- OMDoc
  - Semantics-oriented mathematical representation format
  - Covers the whole range of written mathematics
  - Three levels:
    - Object level – formulae in Content MathML or OpenMath
    - Statement level – definitions, theorems, proofs, examples and the relations between them
    - Theory level – set of contextually related statements, OMDoc theories are compatible to OpenMath content dictionaries

Mathematical search engines

- MathDex
  - Can process semantically poor documents
  - Multiple fields for different mathematical constructs (e.g. numerators, superscripts…)
- LeActiveMath
  - Intelligent, web-based learning system for mathematics
  - Heavily document dependent (own OMDoc with OpenMath)
- LaTeXSearch
  - By Springer publishing
  - Uses LaTeX string representations of formulae
  - Similarity algorithms enable similar formulae matches

Mathematical search engines

- MathDex
  - Can process semantically poor documents
  - Multiple fields for different mathematical constructs (e.g. numerators, superscripts…)
- LeActiveMath
  - Intelligent, web-based learning system for mathematics
  - Heavily document dependent (own OMDoc with OpenMath)
- LaTeXSearch
  - By Springer publishing
  - Uses LaTeX string representations of formulae
  - Similarity algorithms enable similar formulae matches

Mathematical search engines

- MathDex
  - Can process semantically poor documents
  - Multiple fields for different mathematical constructs (e.g. numerators, superscripts…)
- LeActiveMath
  - Intelligent, web-based learning system for mathematics
  - Heavily document dependent (own OMDoc with OpenMath)
- LaTeXSearch
  - By Springer publishing
  - Uses LaTeX string representations of formulae
  - Similarity algorithms enable similar formulae matches

Mathematical search engines II

- MathWebSearch
  - Not based on a full-text engine – substitution trees (known from automatic theorem provers)
  - Document dependant
- EgoMath
  - Mathematical extension to a conventional search engine
  - Aims to search in real-world documents
  - Generalization algorithms

Mathematical search engines II

- MathWebSearch
  - Not based on a full-text engine – substitution trees (known from automatic theorem provers)
  - Document dependant
- EgoMath
  - Mathematical extension to a conventional search engine
  - Aims to search in real-world documents
  - Generalization algorithms

Mathematical search engines comparison

	Input documents	Internal representation	Used converters	Approach	α-eq.	Query language	Queries	Indexing core
MathDex	HTML, TEX/LATEX, Word, PDF	Presentation MathML (text)	jtidy, blahtex, LaTeXML, Hermes, Word+MathType, pdf2tiff->Infty	syntactic	✗	?	text, math, mixed	Apache Lucene
LeActiveMath	OMDoc, OpenMath	OpenMath (text)	-	syntactic	✗	OpenMath (palette editor)	text, math, mixed	Apache Lucene
LATEXSearch	LATEX	LATEX(text)	-	syntactic	✗	LATEX	titles, math, DOI	?
MathWeb Search	Presentation MathML, Content MathML, OpenMath	Content MathML, OpenMath (substitution trees)	-	semantic	✔	QMath, LATEX, Mathematica, Maxima, Maple, Yacas styles (palette editor)	text, math, mixed	Apache Lucene (for text only)
EgoMath	Presentation MathML, Content MathML, PDF	Presentation MathML (text)	Infty	mixed	✗	LATEX	text, math, mixed	EgoThor

Development of a mathematical indexing and searching engine

- Needs:
  - Search
    - Exact mathematical formulae
    - Equal formulae with different notation
    - Similar formulae
    - Subformulae
    - Mixed mathematical-textual
  - Show relevant results
- All in reasonable time

Development of a mathematical indexing and searching engine

- Needs:
  - Search
    - Exact mathematical formulae
    - Equal formulae with different notation
    - Similar formulae
    - Subformulae
    - Mixed mathematical-textual
  - Show relevant results
- All in reasonable time

Development of a mathematical indexing and searching engine

- Needs:
  - Search
    - Exact mathematical formulae
    - Equal formulae with different notation
    - Similar formulae
    - Subformulae
    - Mixed mathematical-textual
  - Show relevant results
- All in reasonable time

Development of a mathematical indexing and searching engine

- Needs:
  - Search
    - Exact mathematical formulae
    - Equal formulae with different notation
    - Similar formulae
    - Subformulae
    - Mixed mathematical-textual
  - Show relevant results
- All in reasonable time

Math Indexer and Searcher

- Indexing phase
  - Mathematical notation recognition
  - Formulae tokenization
    - Presentation MathML – tree structure
  - Modification algorithms – produces N representations for input (sub)formulae
    - Unification of variables
    - Unification of constants
  - Ranking based on
    - Subformula tree depth level
    - Modifications

Math Indexer and Searcher

- Indexing phase
  - Mathematical notation recognition
  - Formulae tokenization
    - Presentation MathML – tree structure
  - Modification algorithms – produces N representations for input (sub)formulae
    - Unification of variables
    - Unification of constants
  - Ranking based on
    - Subformula tree depth level
    - Modifications

Math Indexer and Searcher

- Indexing phase
  - Mathematical notation recognition
  - Formulae tokenization
    - Presentation MathML – tree structure
  - Modification algorithms – produces N representations for input (sub)formulae
    - Unification of variables
    - Unification of constants
  - Ranking based on
    - Subformula tree depth level
    - Modifications

Math Indexer and Searcher

- Indexing phase
  - Mathematical notation recognition
  - Formulae tokenization
    - Presentation MathML – tree structure
  - Modification algorithms – produces N representations for input (sub)formulae
    - Unification of variables
    - Unification of constants
  - Ranking based on
    - Subformula tree depth level
    - Modifications

Math Indexer and Searcher

- Indexing phase
  - Mathematical notation recognition
  - Formulae tokenization
    - Presentation MathML – tree structure
  - Modification algorithms – produces N representations for input (sub)formulae
    - Unification of variables
    - Unification of constants
  - Ranking based on
    - Subformula tree depth level
    - Modifications

Math Indexer and Searcher II

- Searching phase
  - Same modification algorithms
  - Query built from all produced representations using OR boolean operator
  - Reporting relevant results
    - Based on ranking function
- Indexing core
  - Apache Lucene

Math Indexer and Searcher II

- Searching phase
  - Same modification algorithms
  - Query built from all produced representations using OR boolean operator
  - Reporting relevant results
    - Based on ranking function
- Indexing core
  - Apache Lucene

Math Indexer and Searcher II

- Searching phase
  - Same modification algorithms
  - Query built from all produced representations using OR boolean operator
  - Reporting relevant results
    - Based on ranking function
- Indexing core
  - Apache Lucene

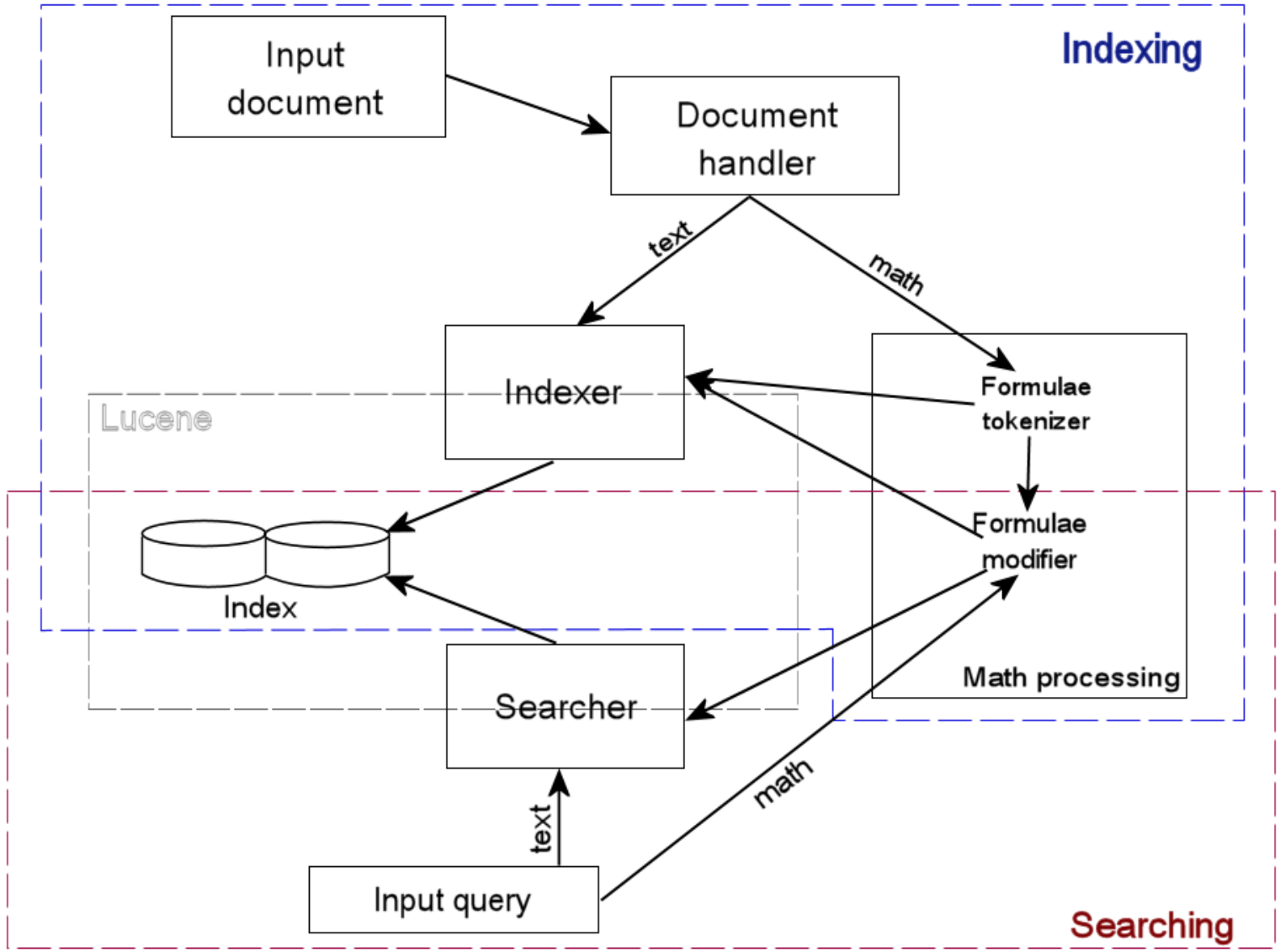
Math Indexer and Searcher II

- Searching phase
  - Same modification algorithms
  - Query built from all produced representations using OR boolean operator
  - Reporting relevant results
    - Based on ranking function
- Indexing core
  - Apache Lucene

Math Indexer and Searcher II

- Searching phase
  - Same modification algorithms
  - Query built from all produced representations using OR boolean operator
  - Reporting relevant results
    - Based on ranking function
- Indexing core
  - Apache Lucene

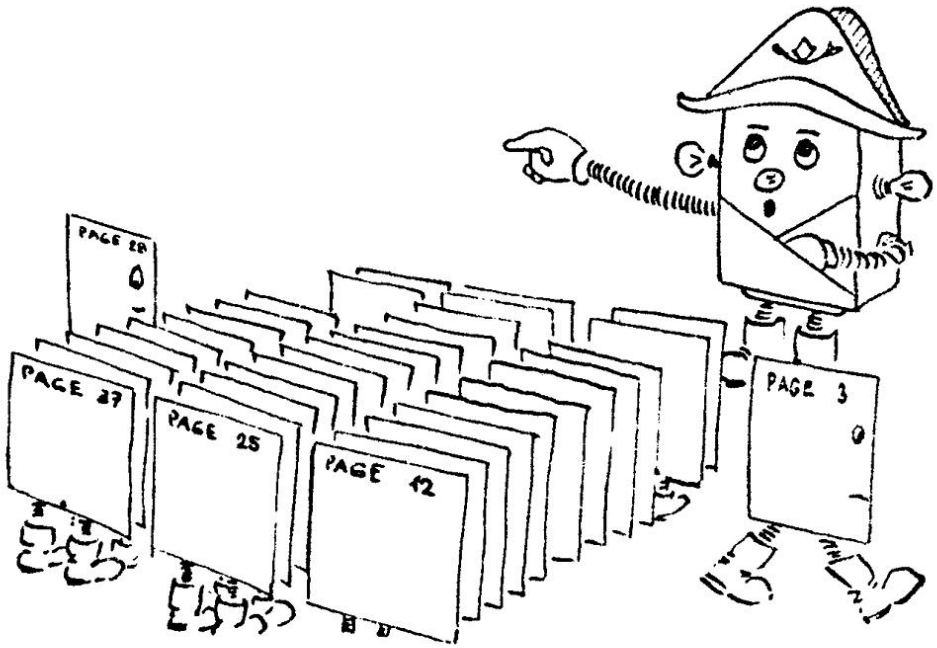
Math Indexer and Searcher design



Conclusion

- Implementation in progress
- Preparation of test corpus of mathematical texts
- Any suggestions welcomed

Questions?



References

Michael Kohlhase and Ioan Sucan:

A Search Engine for Mathematical Formulae.

In Proceedings of Artificial Intelligence and Symbolic Computation, 241–254.

Springer, 2007.

Paul Libbrecht and Erica Melis:

Semantic Search in LeActiveMath.

In Proceedings of the WebALT 2006 Conference.

The WebALT project, 2006.

Jozef Mišutka and Leo Galamboš:

Extending Full Text Search Engine for Mathematical Content.

In DML 2008: Towards Digital Mathematics Library, 55–67.

Masaryk University, 2008.

Robert Miner:

The MathDex Search Engine.

http://www.ima.umn.edu/2006-2007/SW12.8-9.06/activities/Miner-Robert/index.html.

Springer:

LaTeXSearch: About.

http://www.latexsearch.com/LatexTool/static/about.jsp.