Indexing and Searching Mathematics in Digital Libraries

Petr Sojka, Martin Líška

Masaryk University, Faculty of Informatics, Brno, Czech Republic
<sojka@fi.muni.cz> <255768@mail.muni.cz>

20th July, 2011

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Introduction

- Digital mathematics libraries (DML) – access to vast treasure of mathematical papers; 140,000 new papers in Zbl expected this year
- Search – crucial part of MKM; search is a *gate* to this knowledge; DML without math-aware search is an oxymoron
- Text and keyword based search? No problem (Google, referative databases); success
- Mathematics search? It *is* a problem (either in Google or in the referative databases); more or less a failure so far.
- *The* solution has been developed now (not only for EuDML)!

Introduction

- Digital mathematics libraries (DML) – access to vast treasure of mathematical papers; 140,000 new papers in Zbl expected this year
- Search – crucial part of MKM; search is a *gate* to this knowledge; DML without math-aware search is an oxymoron
- Text and keyword based search? No problem (Google, referative databases); success
- Mathematics search? It *is* a problem (either in Google or in the referative databases); more or less a failure so far.
- *The* solution has been developed now (not only for EuDML)!

Introduction

- Digital mathematics libraries (DML) – access to vast treasure of mathematical papers; 140,000 new papers in Zbl expected this year
- Search – crucial part of MKM; search is a *gate* to this knowledge; DML without math-aware search is an oxymoron
- Text and keyword based search? No problem (Google, referative databases); success
- Mathematics search? It *is* a problem (either in Google or in the referative databases); more or less a failure so far.
- *The* solution has been developed now (not only for EuDML)!

Introduction

- Digital mathematics libraries (DML) – access to vast treasure of mathematical papers; 140,000 new papers in Zbl expected this year
- Search – crucial part of MKM; search is a *gate* to this knowledge; DML without math-aware search is an oxymoron
- Text and keyword based search? No problem (Google, referative databases); success
- Mathematics search? It *is* a problem (either in Google or in the referative databases); more or less a failure so far.
- *The* solution has been developed now (not only for EuDML)!

Introduction

- Digital mathematics libraries (DML) – access to vast treasure of mathematical papers; 140,000 new papers in Zbl expected this year
- Search – crucial part of MKM; search is a *gate* to this knowledge; DML without math-aware search is an oxymoron
- Text and keyword based search? No problem (Google, referative databases); success
- Mathematics search? It *is* a problem (either in Google or in the referative databases); more or less a failure so far.
- *The* solution has been developed now (not only for EuDML)!

Search Problem and Existing Approaches

Search problem formulation: given query containing text and formulae, find the most relevant documents.

- MathDex
- EgoMath
- LaTeXSearch
- LeActiveMath
- DLMF equation search
- MathWebSearch

MathDex

Formerly MathFind, seven digit figure NSF grant by Design Science (Robert Miner)

- Lucene based
- Indexing n-grams, presentation MathML
- Multiple fields for different mathematical constructs (e.g. numerators, superscripts)
- Uses several converters and filters to convert to XHTML + MathML – HTML (jtidy), TEX/LATEX (blahtex, LATEXML, Hermes), Word (Word+MathType), PDF (pdf2tiff+Infty)

EgoMath and EgoMath2

- Presented yesterday by Josef Mišutka
- Based on full text core Egothor
- Presentation MathML for indexing
- Formulae augmentation, α-equivalence algorithms and relevance calculation
- As a part, MSE dataset is being developed

LaTeXSearch

- Search engine offered by Springer
- Based on searching over TeX math string representations
- Some kind of similarity matching
- Not open source, only for LaTeX math string approximate match, no formulae structure matching
- Small database: 3 million formulae from ‘random’ sources

LeActiveMath

- Indexing string tokens from OMDoc with OpenMath semantic notation
- *Only* for documents authored for LeActiveMath learning environment
- Lucene based

DLMF

- Equation search
- *Only* for documents authored for DLMF in special markup
- Lucene based

MathWeb Search

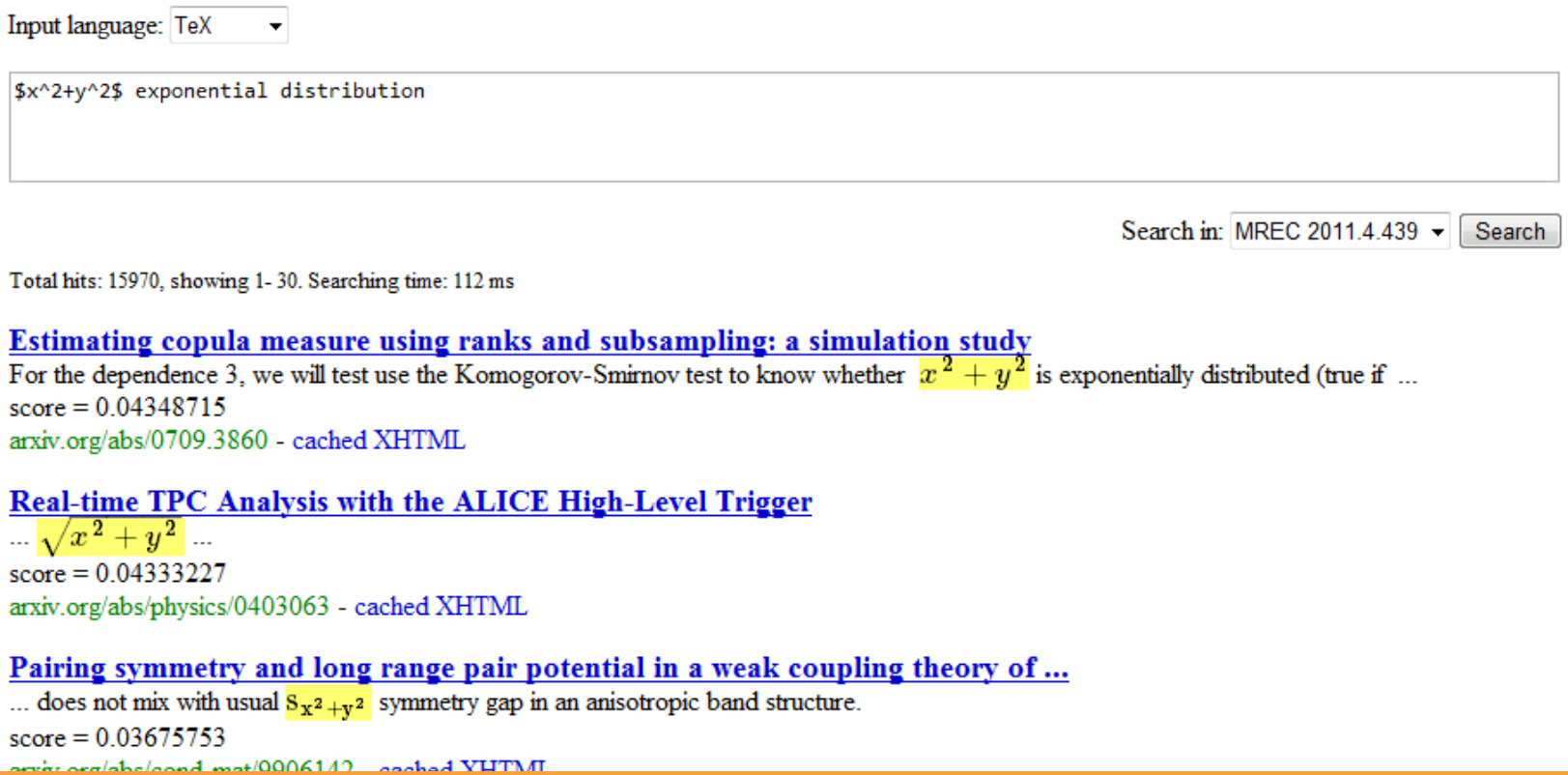
- Not based on full text searching
- Supports Content MathML and OpenMath
- Uses substitution trees – semantic approach

Comparison

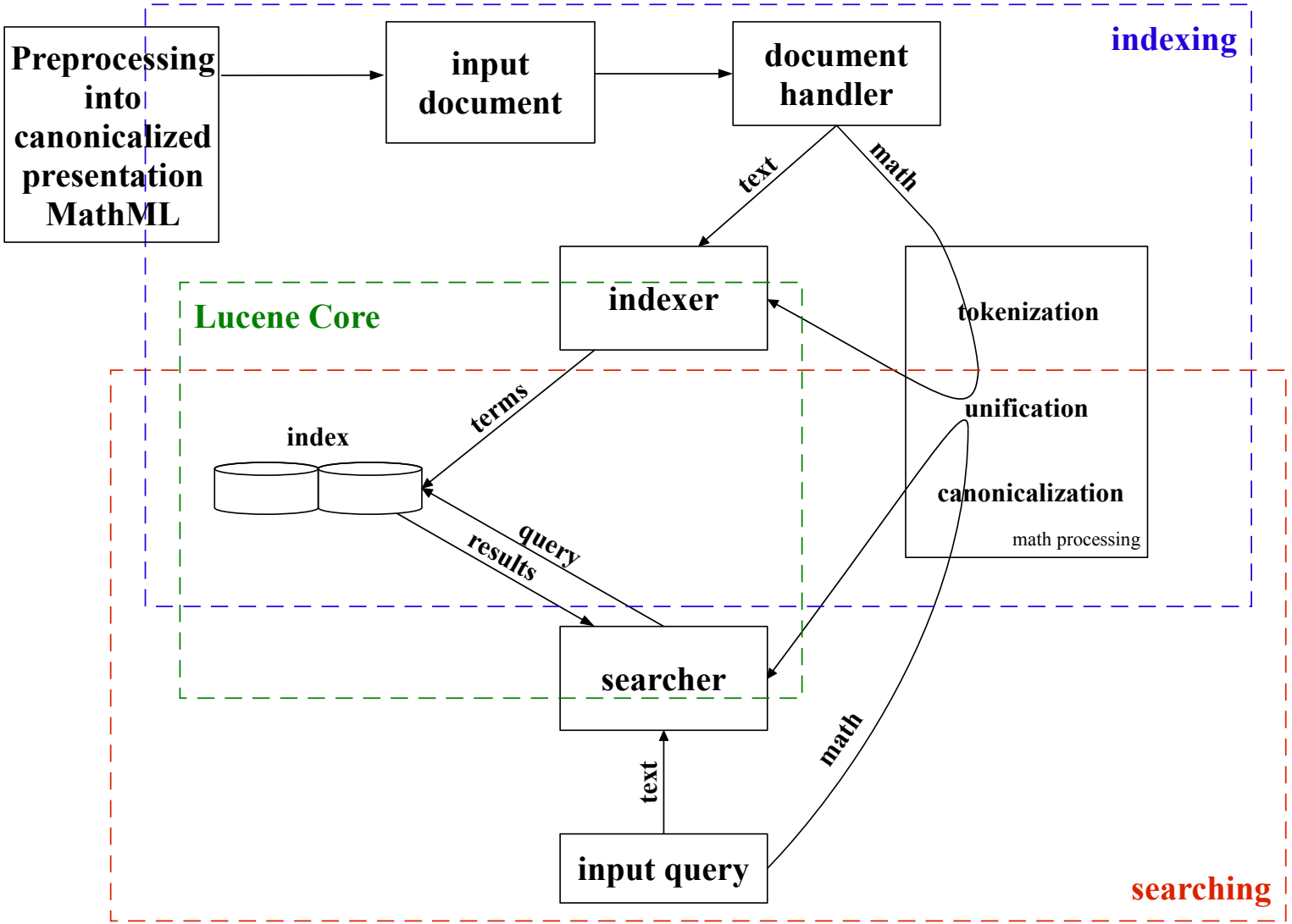
System	Input docs	Internal repres.	Approach	α-eq.	Query language	Queries	Indexing core
MathDex	HTML, TEX/LATEX, Word, PDF	Presentation MathML (as strings)	syntactic	×	?	text, math, mixed	Apache Lucene
LeActiveMath	OMDoc, OpenMath	OpenMath (as string)	syntactic	×	OpenMath (palette editor)	text, math, mixed	Apache Lucene
LATEXSearch	LATEX	LATEX (as string)	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 trees (as strings)	mixed	✔	LATEX	text, math, mixed	EgoThor
MIaS	any (well-formed) MathML	Canonical Presentation MathML trees (as compacted strings)	math tree similarity/ normalization	✔	AMSLaTeX or MathML	text, math, mixed	Apache Lucene/ Solr

Math Indexer and Searcher

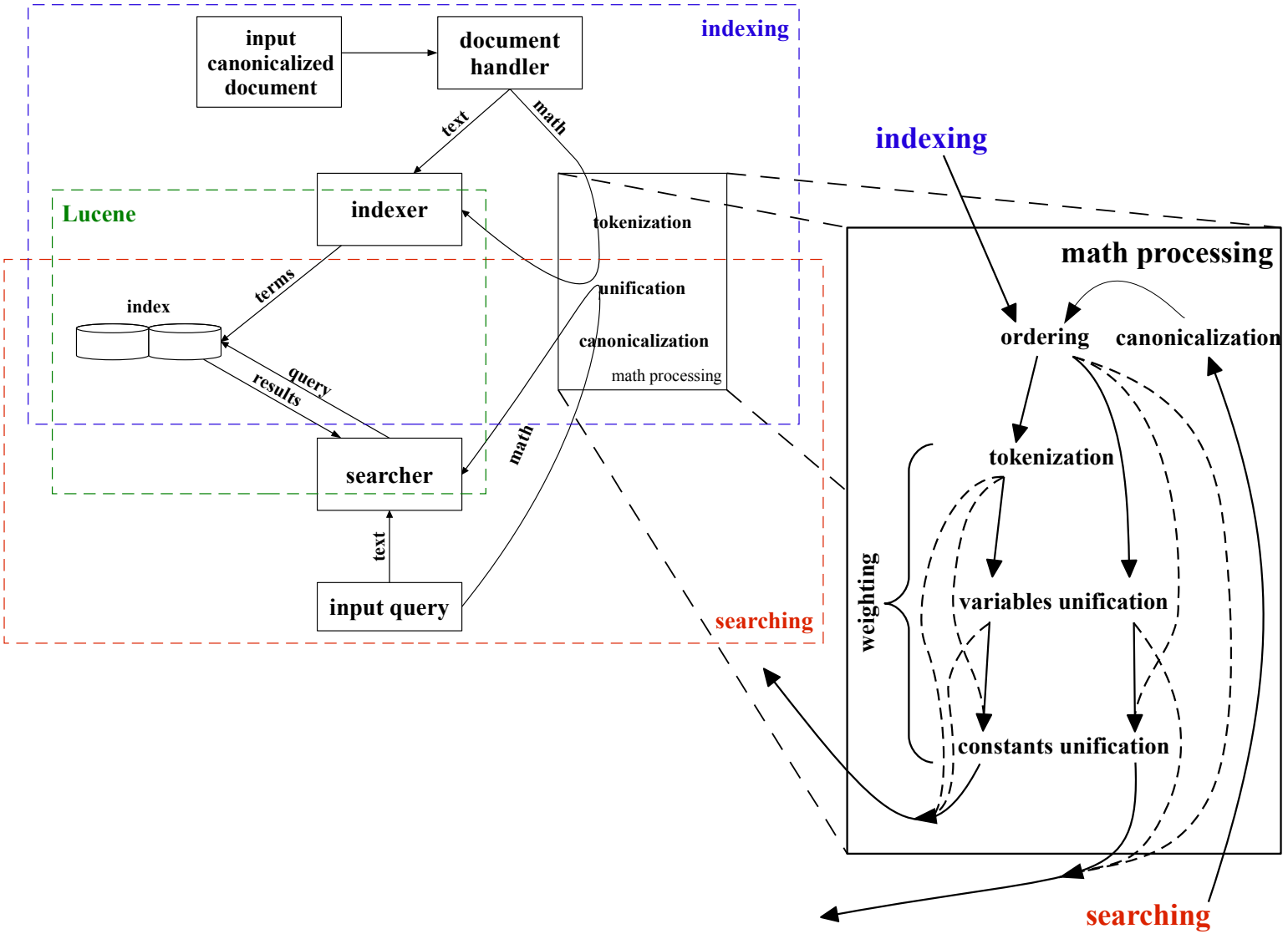
- Math-aware, full-text based search engine
- Joins textual and mathematical querying
- MathML and TEX input



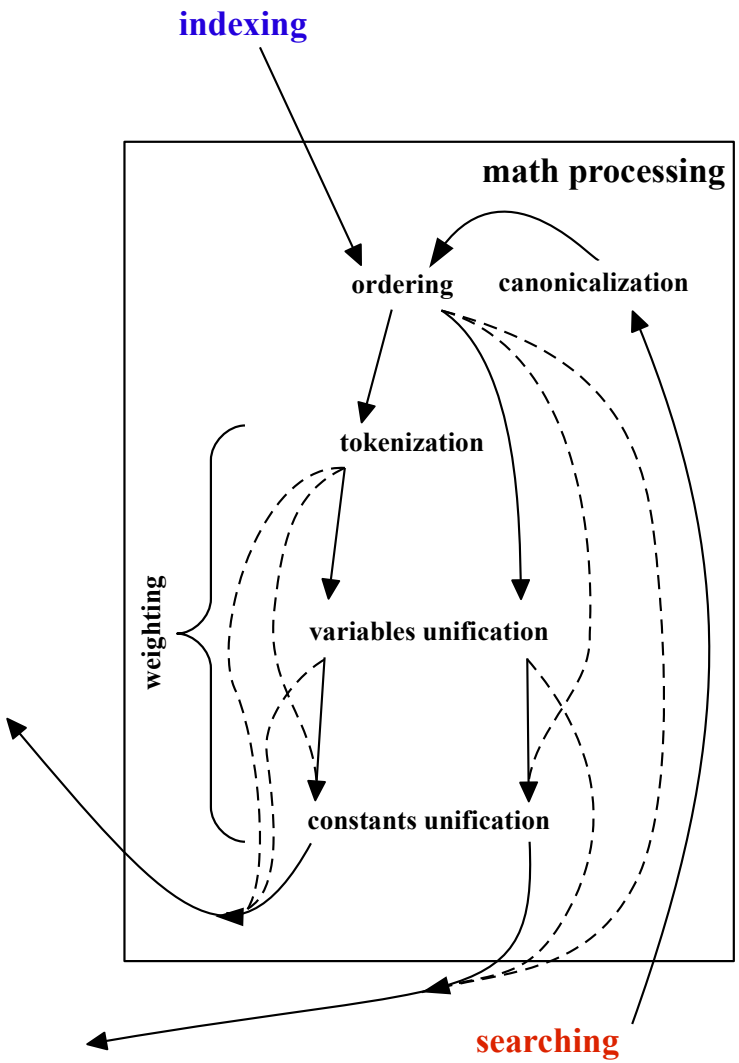
Design


Preprocessing into canonicalized presentation MathML
input document
document handler
indexing
text
math
indexer
Lucene Core
tokenization
unification
canonicalization
math processing
terms
index
query
results
searcher
text
math
input query
searching

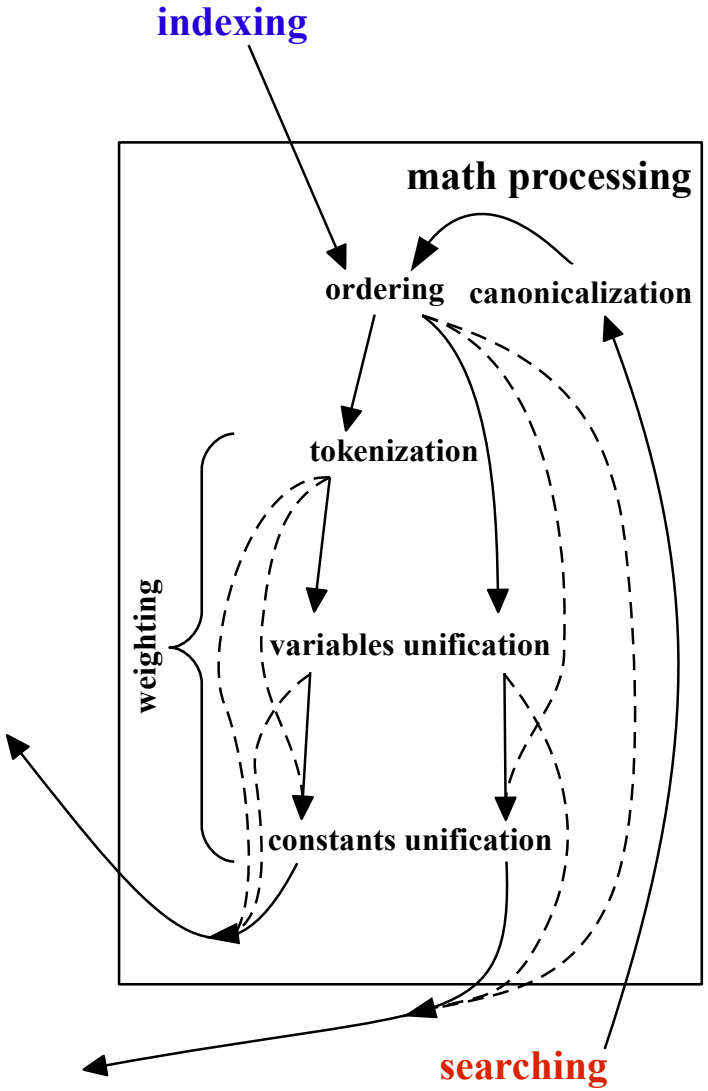
Design II


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

Design III


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

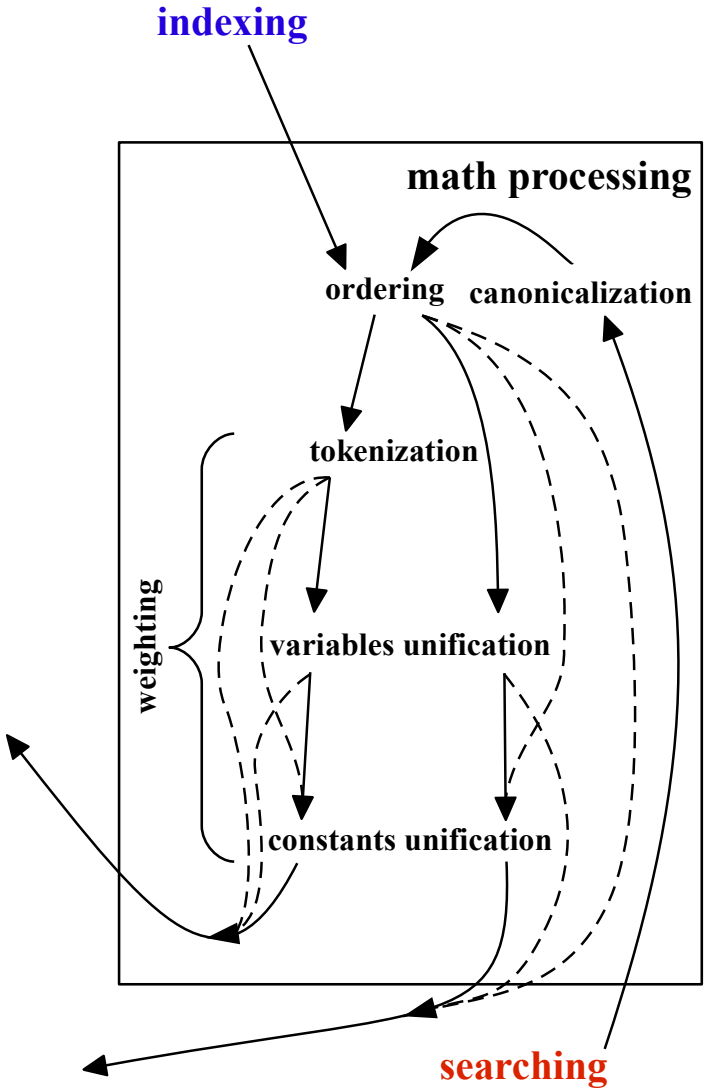
Example



indexing

$x^y + y^3$

Example



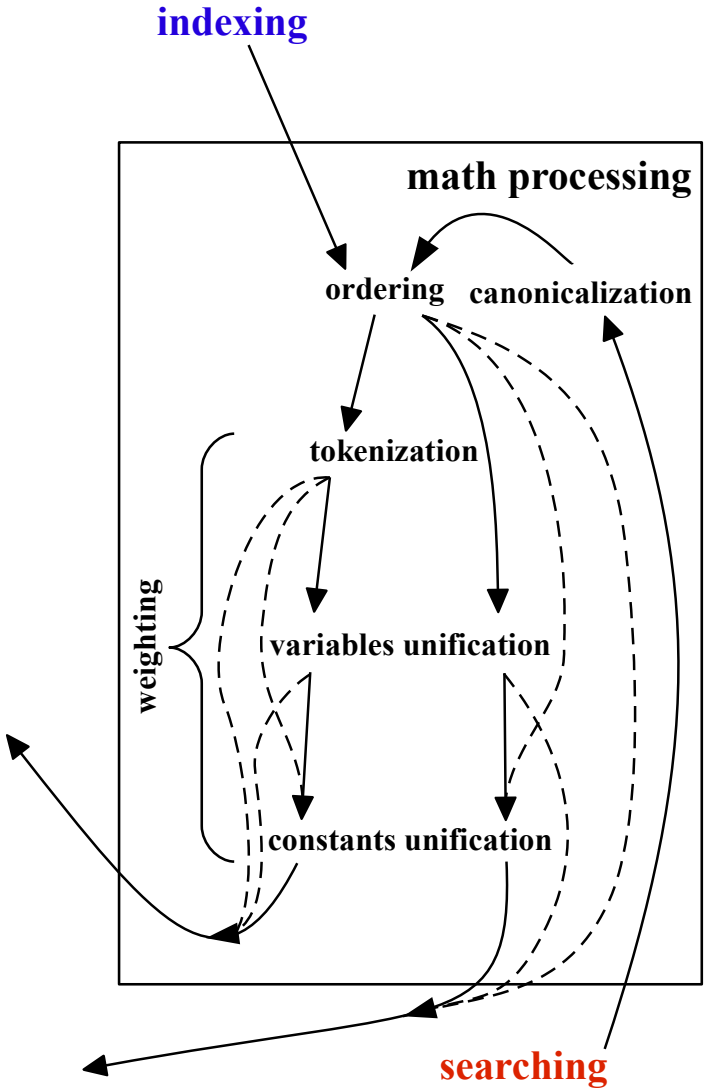
indexing

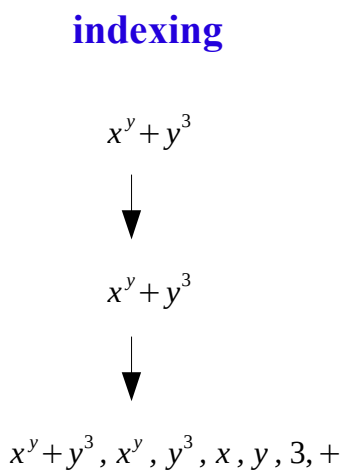
$x^y + y^3$

↓

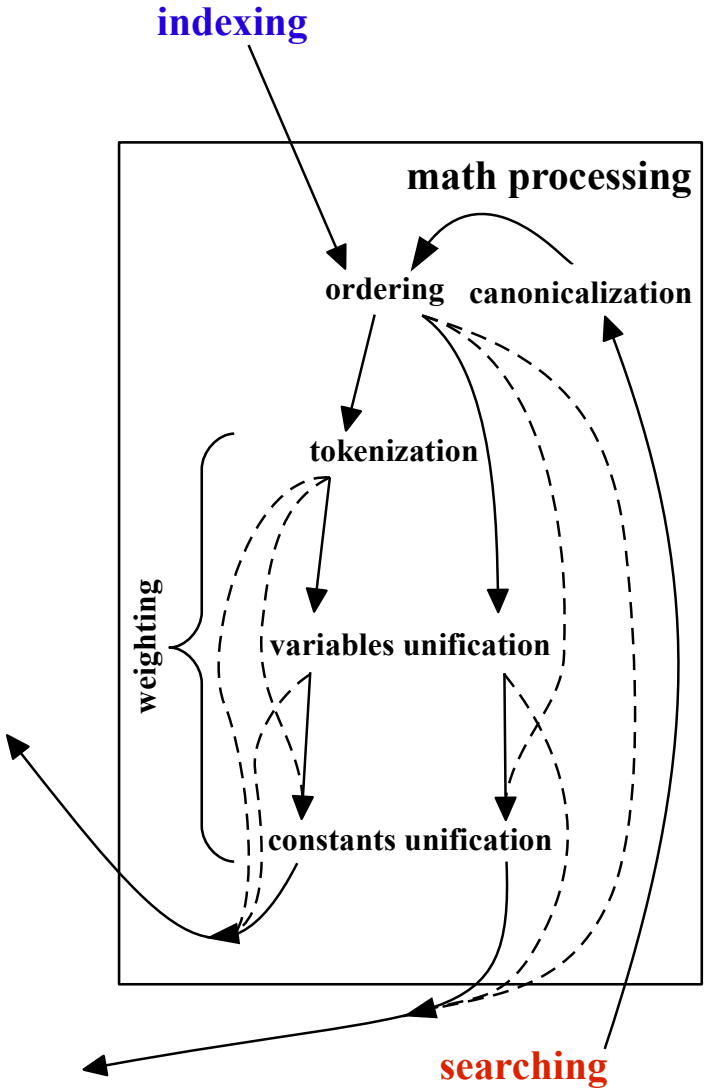
$x^y + y^3$

Example


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


indexing
x^y+y^3
x^y+y^3
$x^y+y^3, x^y, y^3, x, y, 3, +$

Example



indexing

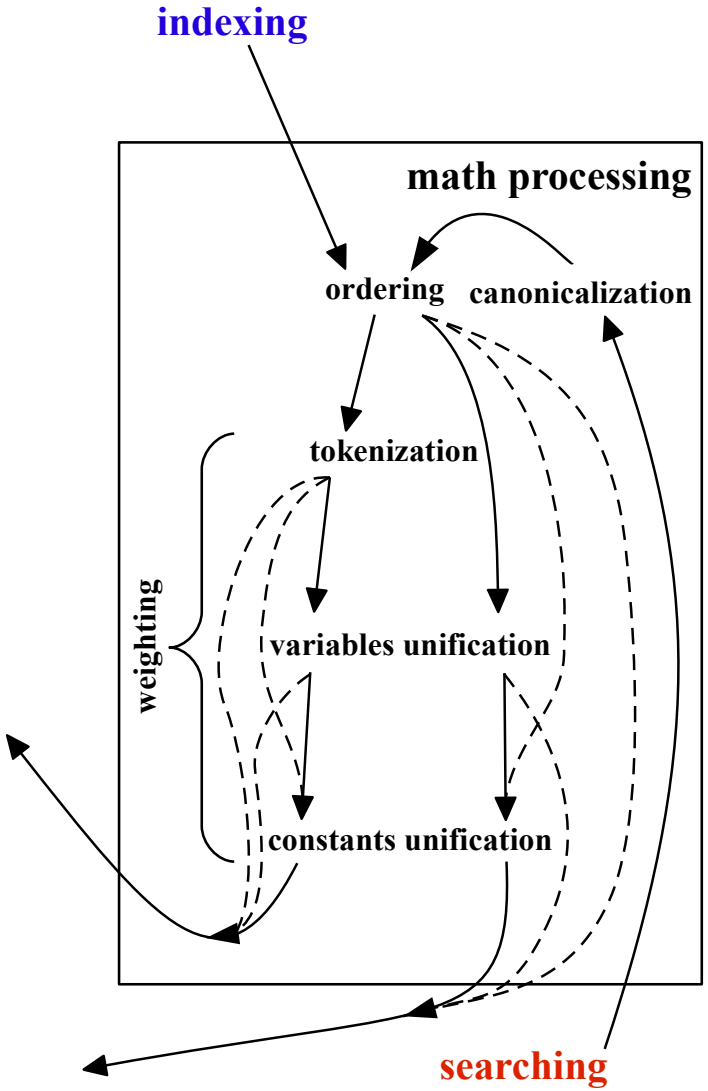
$x^y + y^3$

$x^y + y^3$

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

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

Example



indexing

$$x^y+y^3$$

$$\downarrow$$

$$x^y+y^3$$

$$\downarrow$$

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

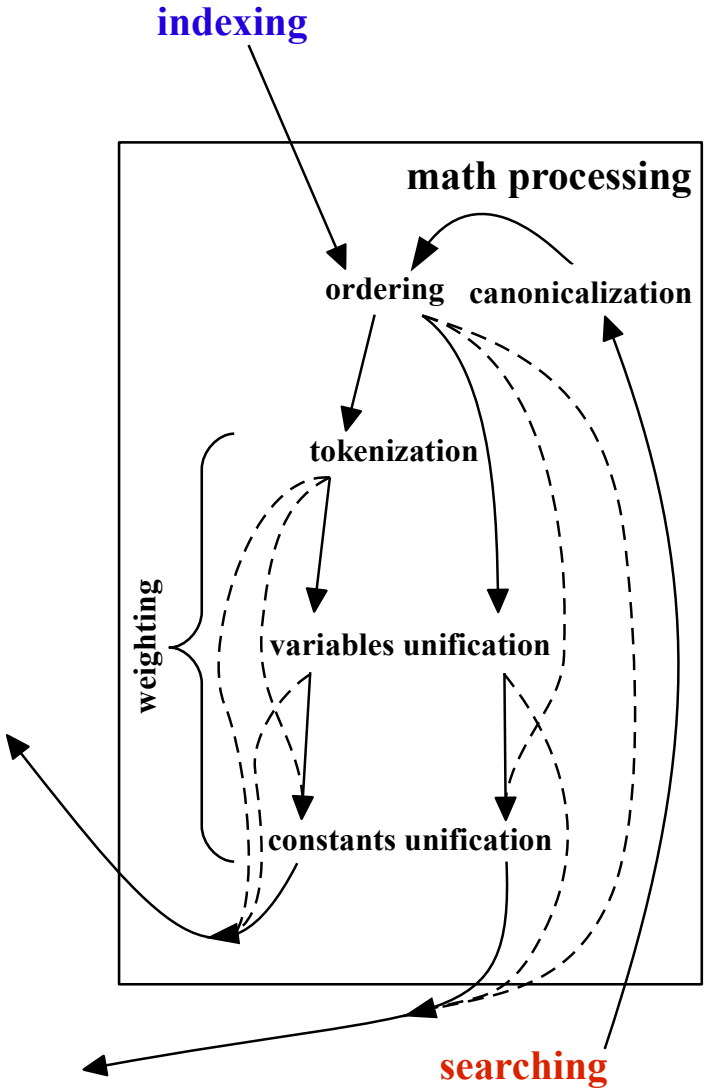
$$\downarrow$$

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

$$\downarrow$$

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

Example



indexing

x^y+y^3

x^y+y^3

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

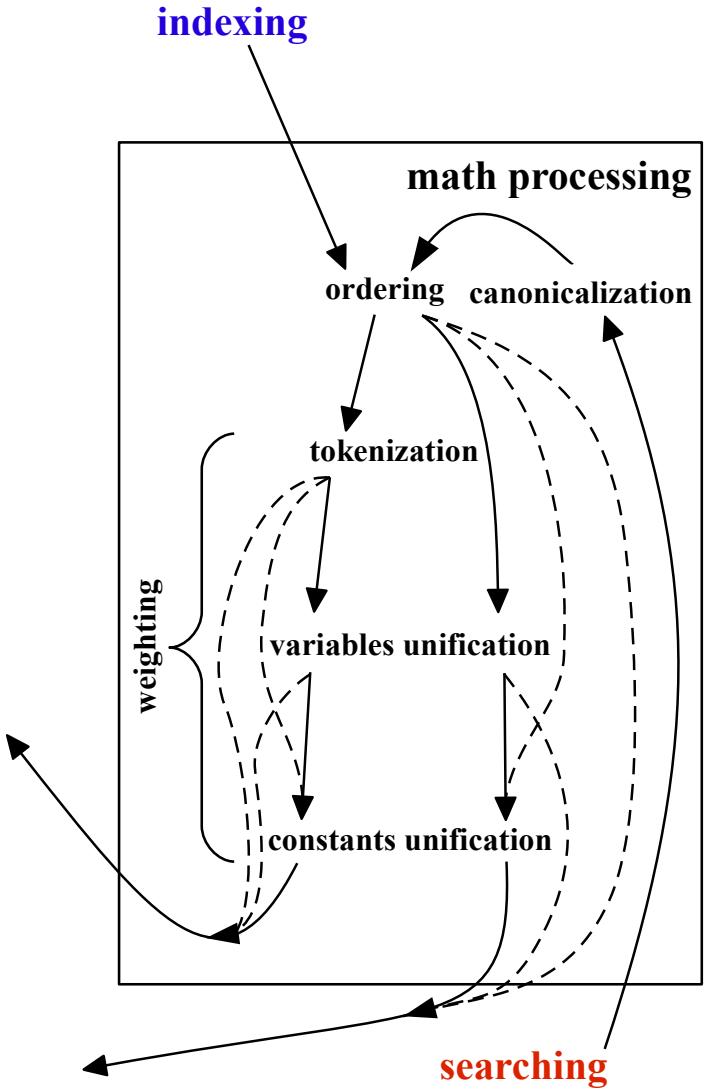
$x^y+y^3\ ,\ x^y\ ,\ y^3\ ,\ x\ ,\ y\ ,\ 3,+\ ,\ id_1^{id_2}+id_2^3,\ id_1^{id_2},\ id_1^3$

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

searching

x^y+y^2

Example



indexing

x^y+y^3

x^y+y^3

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

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

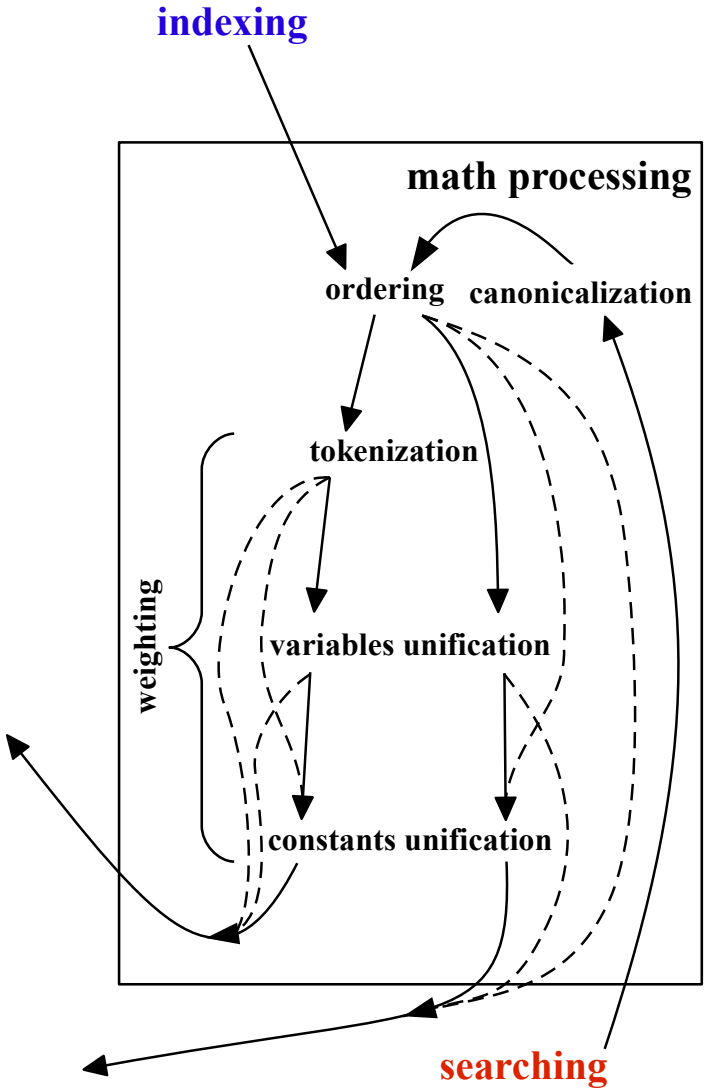
$x^y+y^3\,,\,x^y\,,\,y^3\,,\,x\,,\,y\,,\,3,+\,,\,id_1^{id_2}+id_2^3\,,$
$id_1^{id_2}\,,\,id_1^3\,,\,x^y+y^{const}\,,\,y^{const}\,,\,id_1^{id_2}+id_2^{const}\,,\,id_1^{const}$

searching

x^y+y^2

x^y+y^2

Example



indexing

x^y+y^3

x^y+y^3

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

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

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

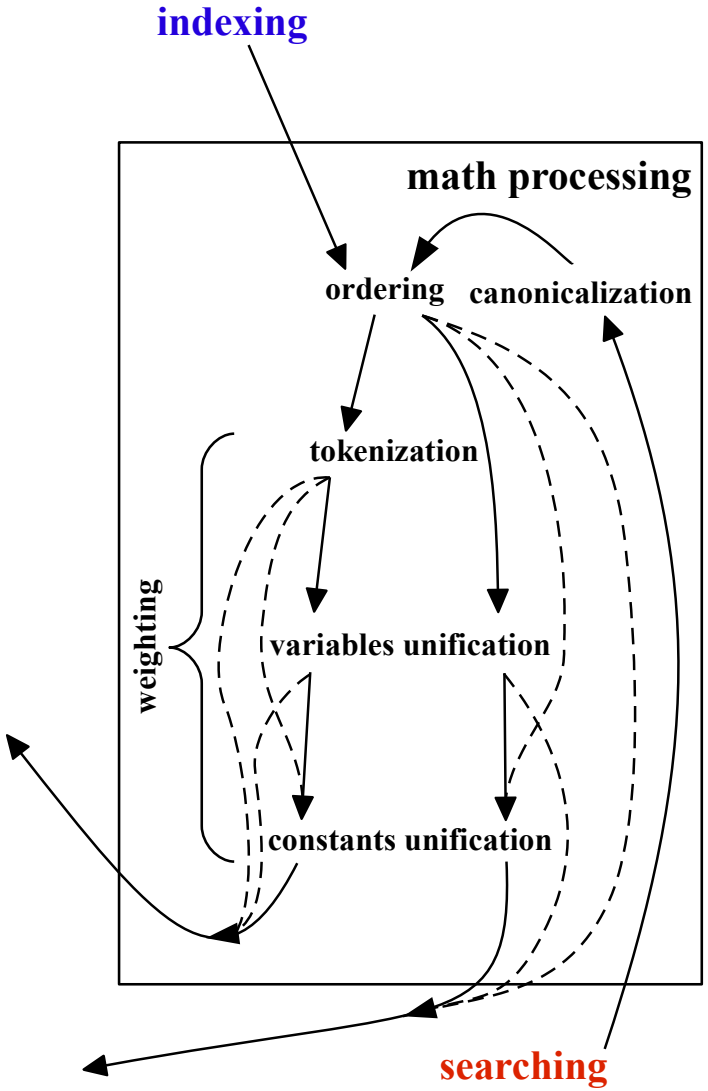
searching

x^y+y^2

x^y+y^2

$x^y+y^2\,,\,id_1^{id_2}+id_2^2$

Example



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

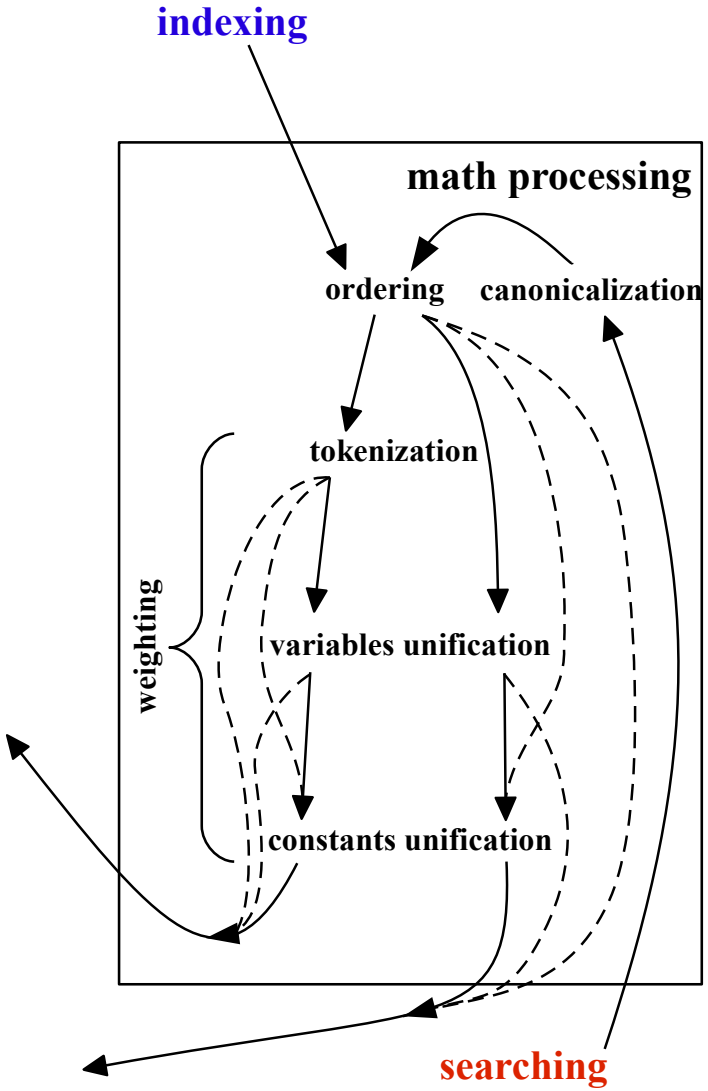
x^y+y^2

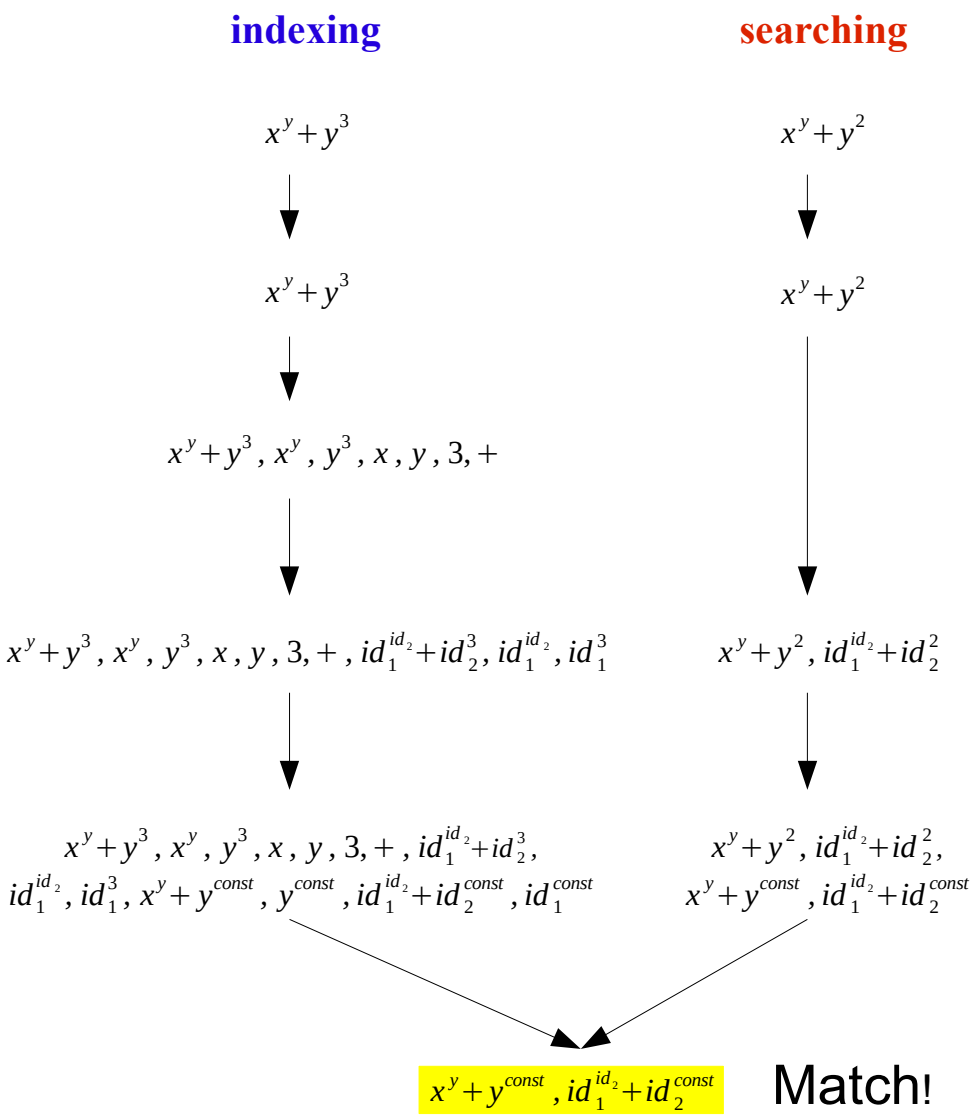
x^y+y^2

$x^y+y^2\,,\ id_1^{id_2}+id_2^2$

$x^y+y^2\,,\ id_1^{id_2}+id_2^2,$
$x^y+y^{const}\,,\ id_1^{id_2}+id_2^{const}$

Example


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


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

Weighting

- We used a weighting utility
- Indexing
  - initial weight of whole formula $= \frac{1}{number\_of\_nodes}$
  - tokenization – level coefficient $l = 0.7$
  - variables unification – coefficient $v = 0.8$
  - number constants unification – coefficient $c = 0.5$
- Searching
  - $result * number\_of\_query\_nodes$

Formula Processing Example

input: $(a+b^{2+c}, 0.125)$ $0.125 = \frac{1}{8} = \textit{formula tree nodes}$

Formula Processing Example

input: $(a+b^{2+c}, 0.125)$

$\Big(\text{“mi”} < \text{“mn”} \Rightarrow 2 <\text{->} c\Big)$

ordering: $(a+b^{c+2}, 0.125)$

Formula Processing Example

input: $(a+b^{2+c}, 0.125)$

ordering: $(a+b^{c+2}, 0.125)$

tokenization: $(a, 0.0875)$ $(+, 0.0875)$ $(b^{c+2}, 0.0875)$

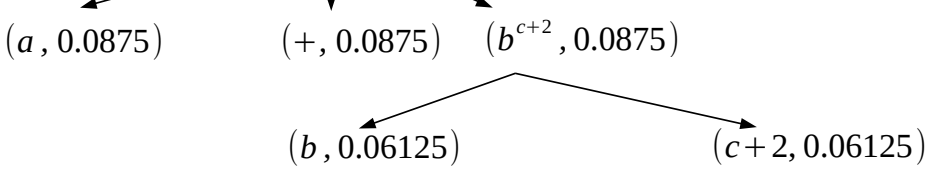
$$0.0875 = 0.125 \cdot 0.7(l)$$

Formula Processing Example

input: $(a+b^{2+c}, 0.125)$

ordering: $(a+b^{c+2}, 0.125)$

tokenization:



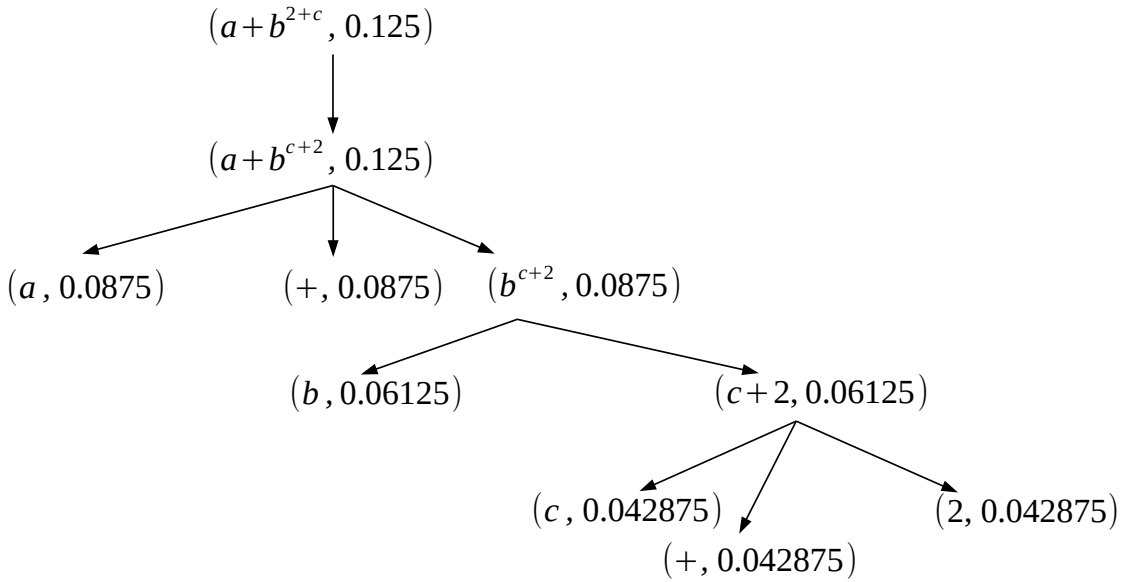
$$0.06125 = 0.0875 \cdot 0.7(l)$$

Formula Processing Example

input:

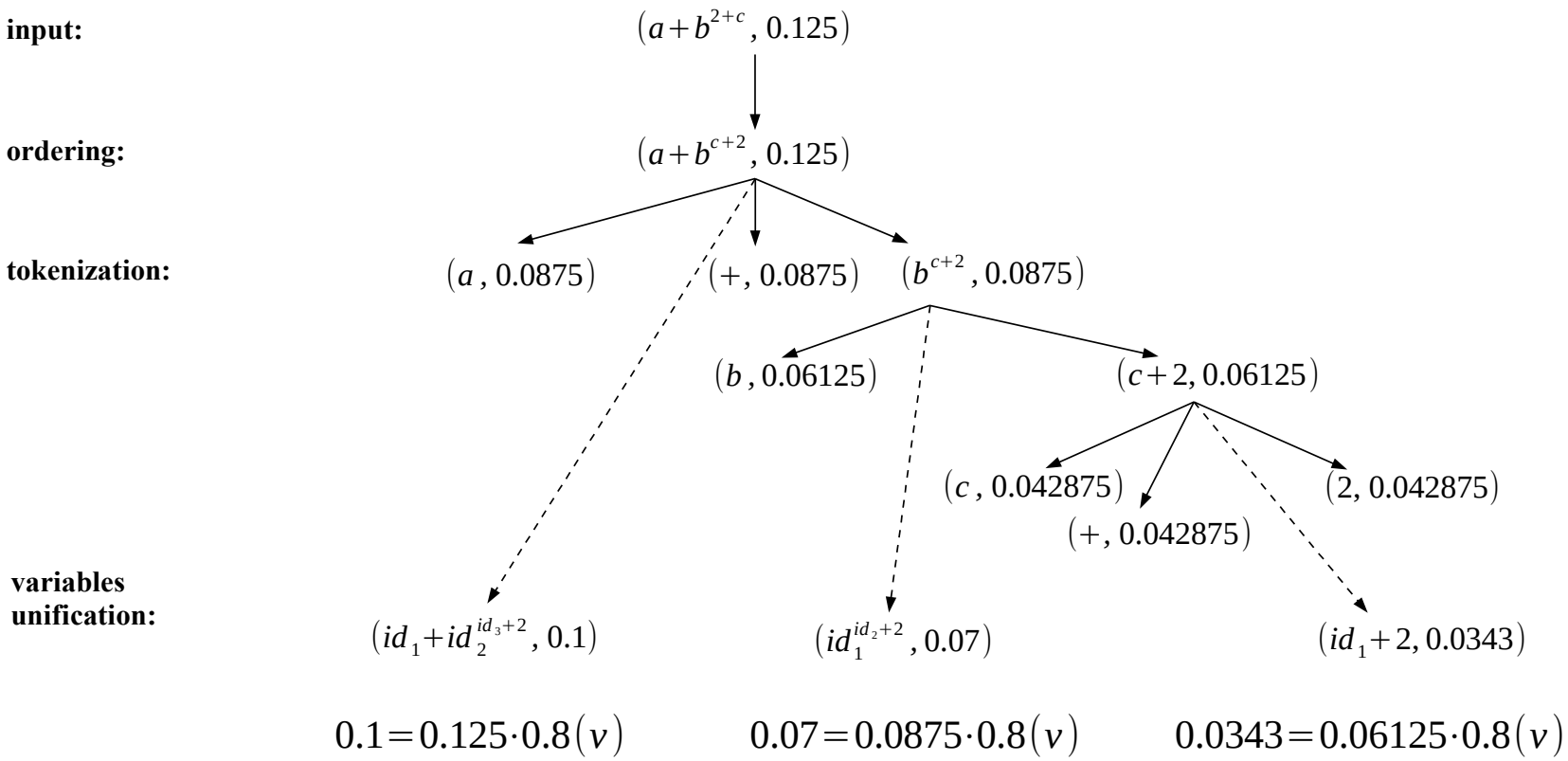
ordering:

tokenization:

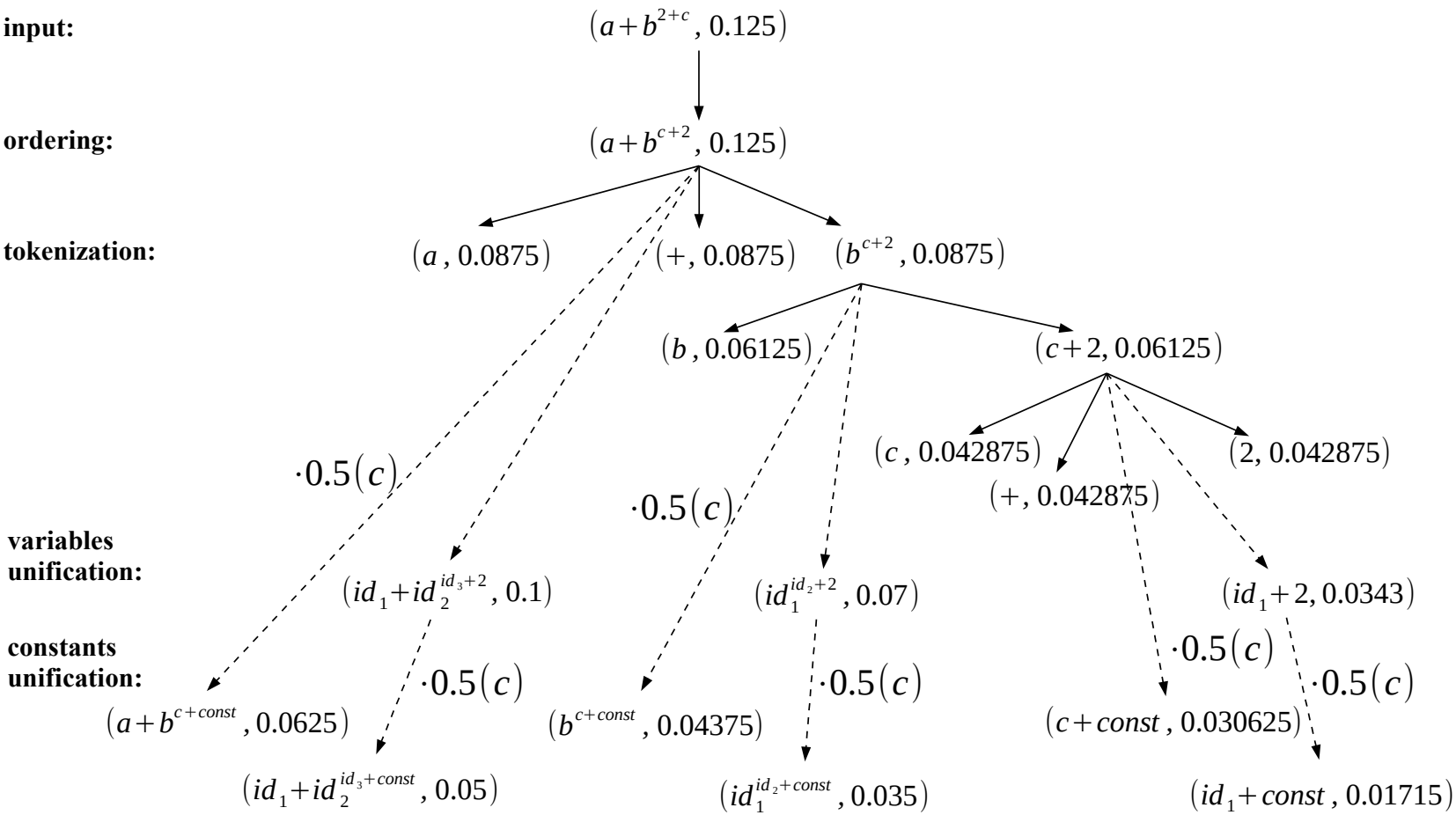


$$0.042875 = 0.06125 \cdot 0.7(l)$$

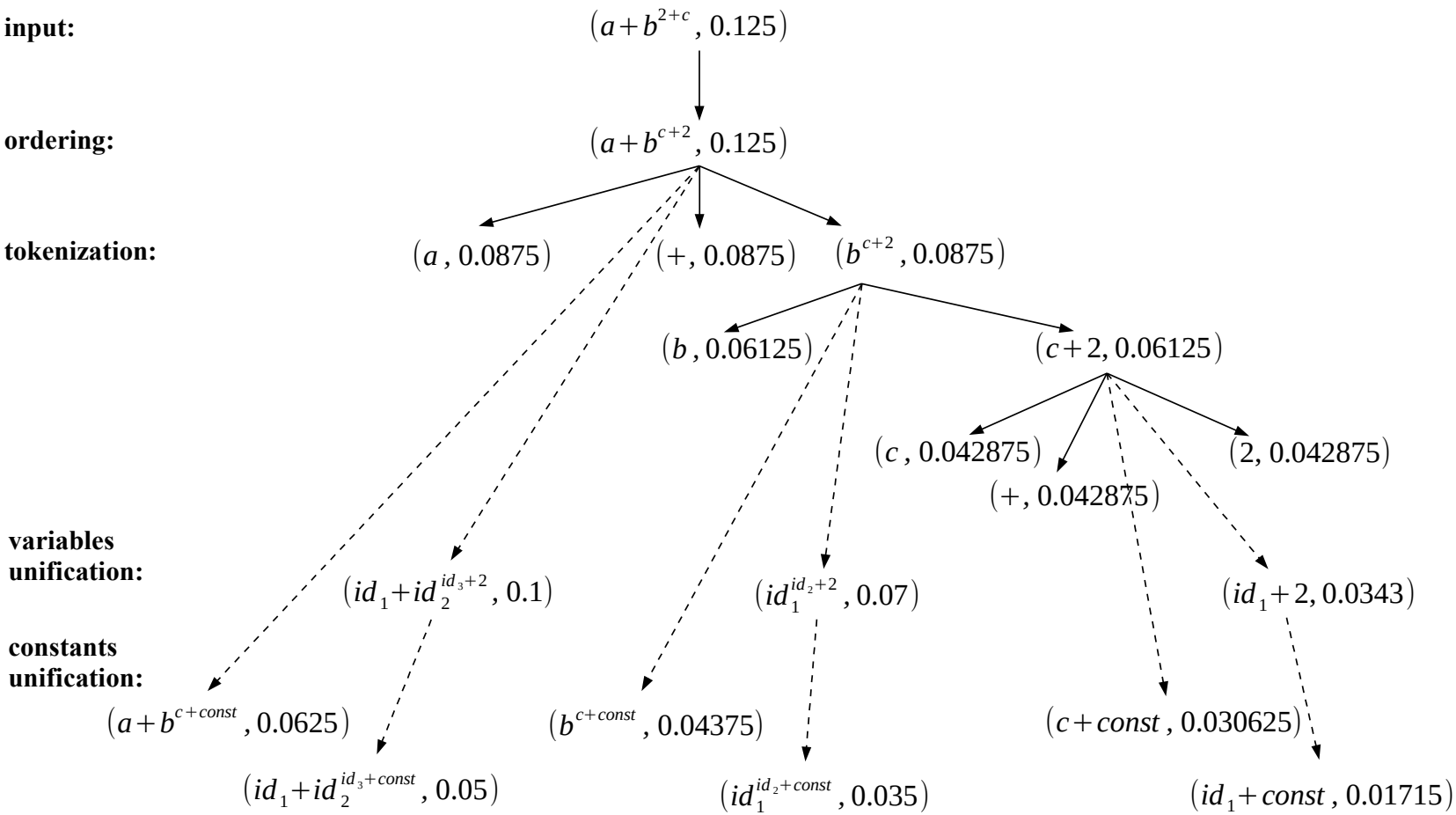
Formula Processing Example



Formula Processing Example



Formula Processing Example



Implementation

- Java
- Lucene 3.1.0
- Mathematical part implements Lucene's interface Tokenizer – able to integrate to any Lucene based system
  - MIaS4Solr plugin was created for the use in Solr in EuDML
- Textual content – processed by StandardAnalyzer

MREC

- MREC 2011.4.439
  - 439,423 documents (originated from arXMLiv, validated, enriched with metadata for snippet generation)
  - Uncompressed size 124 GB, compressed 15 GB
  - 158 million input formulae, 2.9 billion expressions indexed
- For more information see [1] and visit the DML presentation about MREC and WebMIaS tomorrow at 9:30AM

Scalability

- MREC 2011.4.439
  - Indexing time: 1,378.82 min (23 hours)
  - Average query time: 469 ms
  - Index size 63 GB
  - Linear time scale – still seems feasible for a digital library

WebMIaS


Examples About Help Contact
WEBMIaS
MATH INDEXER AND SEARCHER
Input language: MathML
<math><mrow><msup><mi>x</mi> <mn>2</mn> </msup><mo>+</mo><msup><mi>y</mi> <mn>2</mn> </msup></mrow></math>
Canonicalized MathML query:
<math xmlns="http://www.w3.org/1998/Math/MathML">
<mrow>
<msup> <mi>x</mi><mn>2</mn></msup>
<mo>+</mo>
<msup> <mi>y</mi><mn>2</mn></msup>
</mrow>
</math>
Search in: MREC 2011.4.439 Search
Total hits: 36817, showing 1- 30. Searching time: 100 ms
Finite Precision Measurement Nullifies Euclid's Postulates
... and the unit circle $x^2 + y^2 = 1$ are both dense but they do not intersect, in contradiction to Euclid's postulates ...
score = 0.19934596
arxiv.org/abs/quant-ph/0310035 - cached XHTML
COMMENT ON RECENT TUNNELING MEASUREMENTS ON Bi22Sr22CaCu22O88
... gap, (b) s-wave gap, and (c) $s_{x^2+y^2}$ gap.

WebMIaS

- Demo web interface: http://nlp.fi.muni.cz/projekty/eudml/mias
  - MathML/TEX input (Tralics for conversion to MathML)
  - Canonicalization of the query – UMCL
  - Matched document snippet generation
  - MathJax for nicer math rendering and better portability
- DML presentation about MREC and WebMIaS tomorrow at 09:30

Conclusion

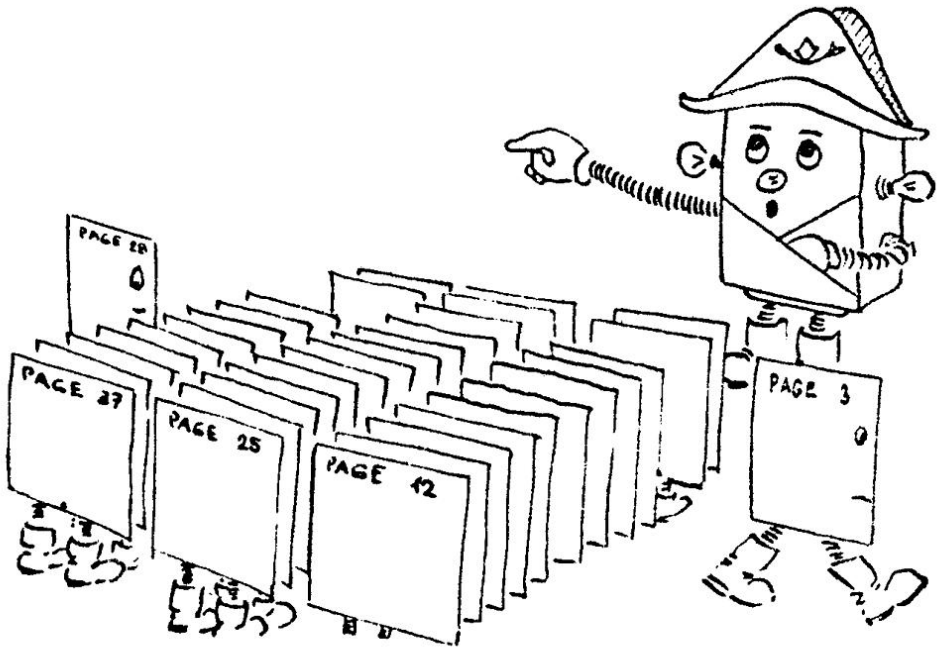
- Project pages – http://nlp.fi.muni.cz/projekty/eudml/mias
- Future work
  - Design a more complex ordering algorithm
  - Improved MathML canonicalization and new preprocessing filters, employment and testing on EuDML data
  - Weighting optimization
  - Query relaxation (“have you meant…”)
  - ?Addition of Content MathML tree indexing?
  - ?Mathematical equivalence computation via symbolic algebra system?
  - ?Cooperating with other MSE systems developers to have *the* a *final* solution of MSE for DML and MKM communities.

Conclusion

MIaS is *the* MSE

- *text+math IR compatible* (fits mathematician's needs)
- new math similarity (weighting) approach compatible with both presentation and content MathML
- *scalable* (index with almost 3 billion formulae tested)
- *Lucene/Solr compatible* system to be employed in EuDML

Questions?




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://www.fi.muni.cz/ sojka/dml-2011-program.html>.


Petr Sojka, editor.
Towards a Digital Mathematics Library, Paris, France, July 2010. Masaryk University.
<http://www.fi.muni.cz/ sojka/dml-2010-program.html>.


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 [2], pp. 105–117, <http://dml.cz/dmlcz/702579>


Kováčik, O., Rákosník, J.: On spaces $L^{p(x)}$ and $W^{k,p(x)}$. Czechoslovak Mathematical Journal 41, 592–618 (1991), <http://dml.cz/dmlcz/102493>


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>


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 [2], 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>.