MIaS – Math Indexer and Searcher

Martin Líška, Petr Sojka, Michal Růžička

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

8th July, 2012

Introduction

- Tool to search mathematics in DMLs
- Search in real world documents
- Easily usable and straightforward

Introduction

- Tool to search mathematics in DMLs
- Search in real world documents
- Easily usable and straightforward

Introduction

- Tool to search mathematics in DMLs
- Search in real world documents
- Easily usable and straightforward

Introduction

- Tool to search mathematics in DMLs
- Search in real world documents
- Easily usable and straightforward

Introduction

- Tool to search mathematics in DMLs

✓ Built on a conventional full text indexing core – capable of indexing millions of documents

- Search in real world documents

- Easily usable and straightforward

Introduction

- Tool to search mathematics in DMLs

✓ Built on a conventional full text indexing core – capable of indexing millions of documents

- Search in real world documents

✓ Uses Presentation MathML for indexing

- Easily usable and straightforward

Introduction

- Tool to search mathematics in DMLs

✓ Built on a conventional full text indexing core – capable of indexing millions of documents

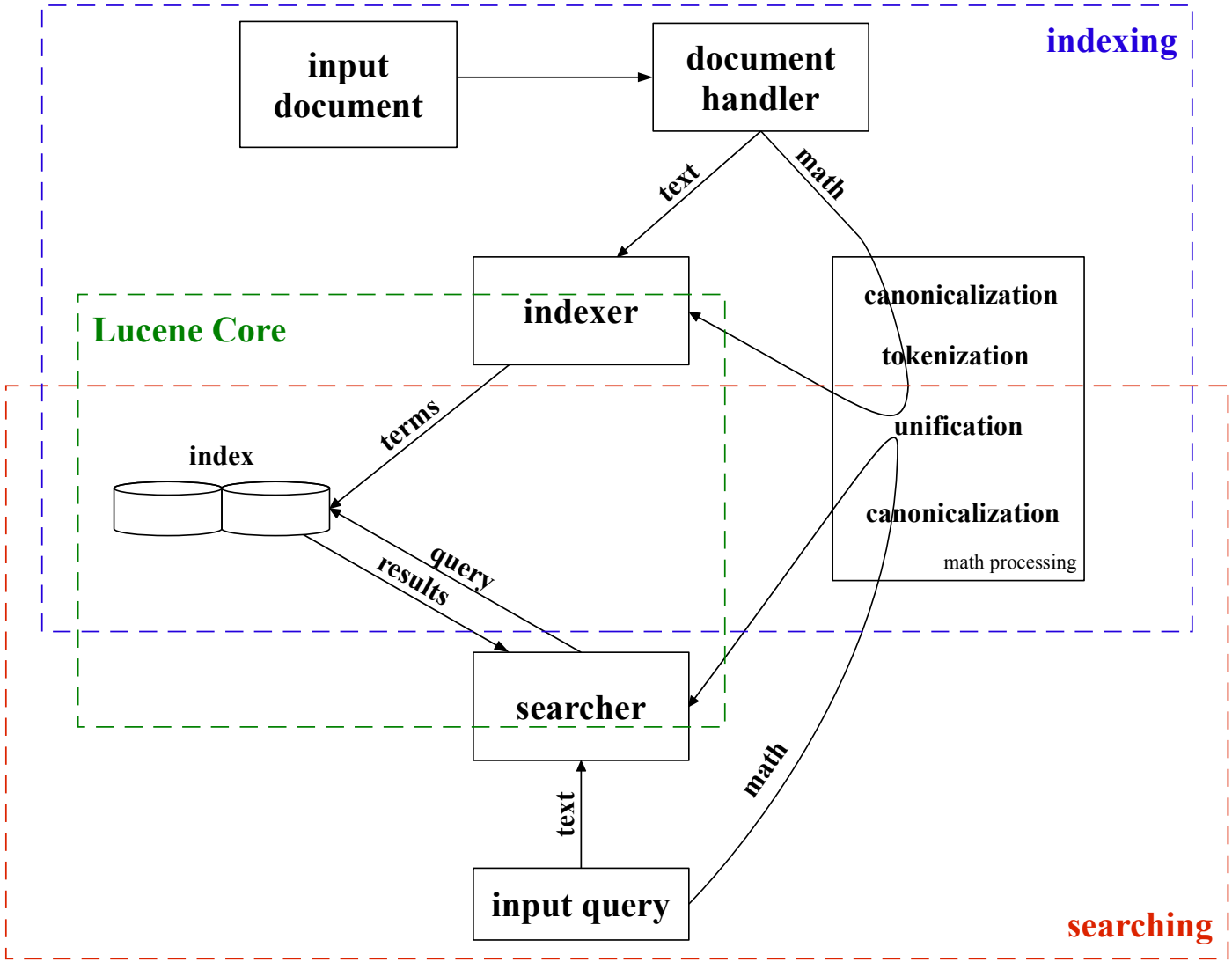
- Search in real world documents

✓ Uses Presentation MathML for indexing

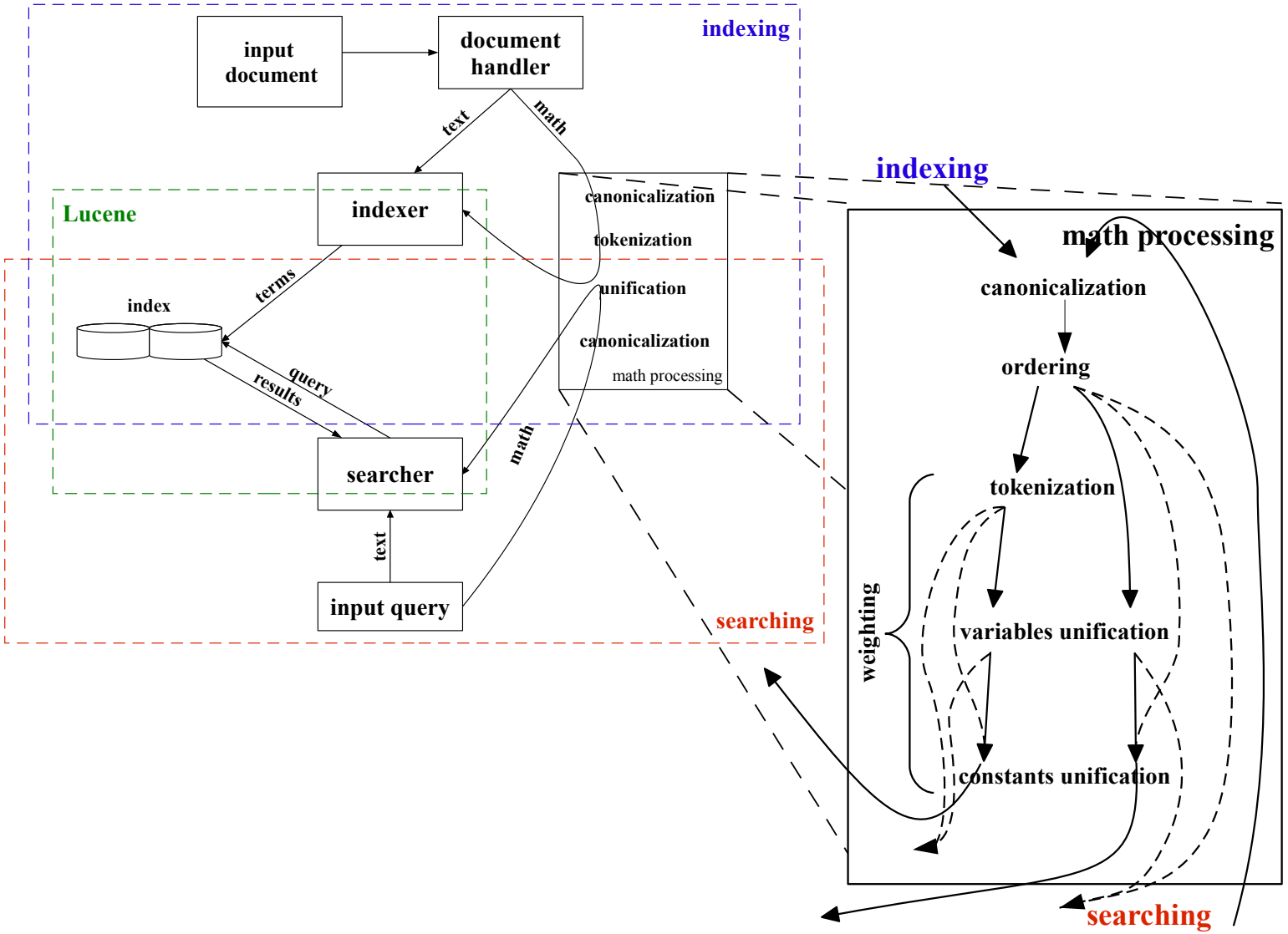
- Easily usable and straightforward

✓ Supports mixed math-textual queries with MathML/TEX input

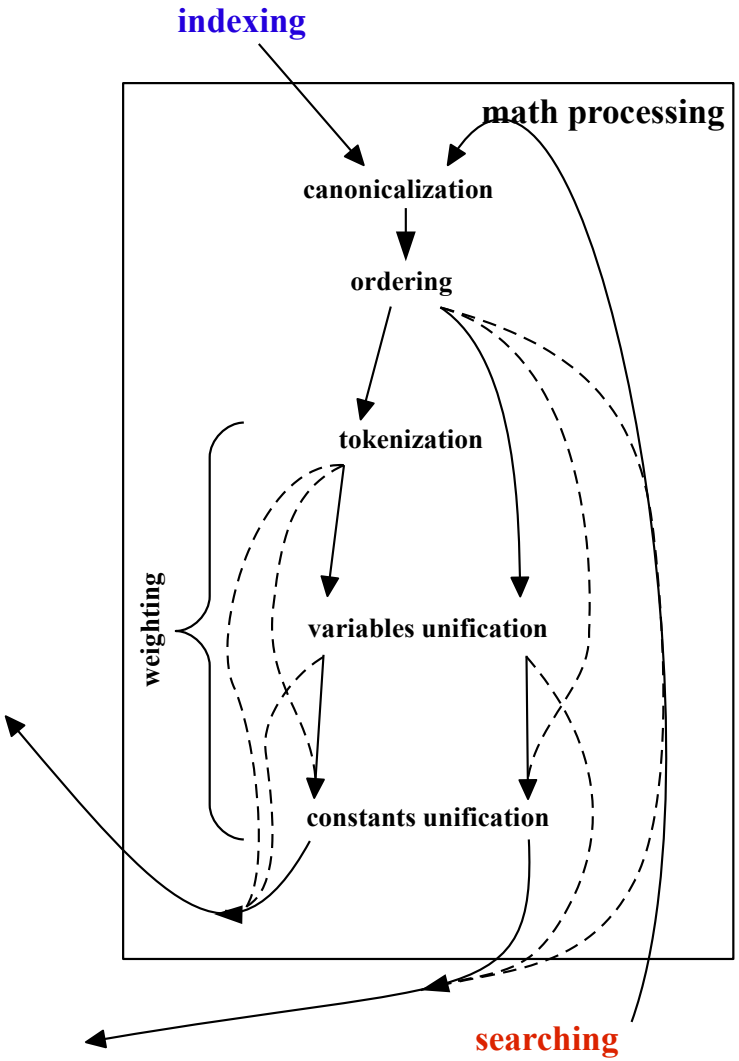
Design


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

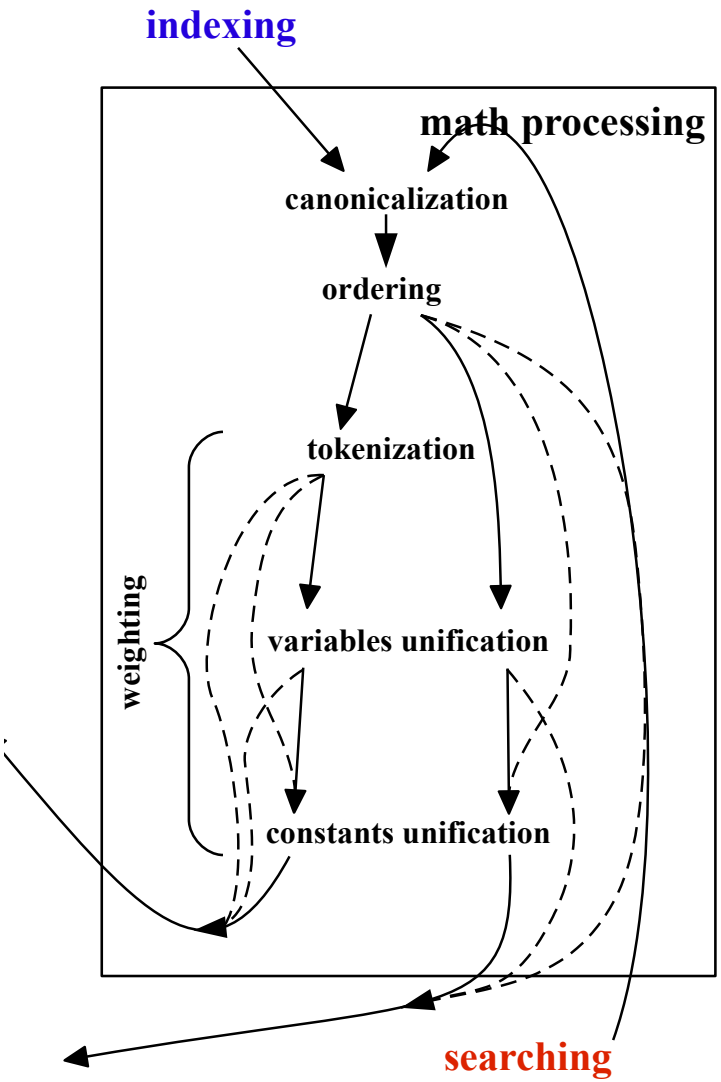
Design II


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

Design III


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

Example

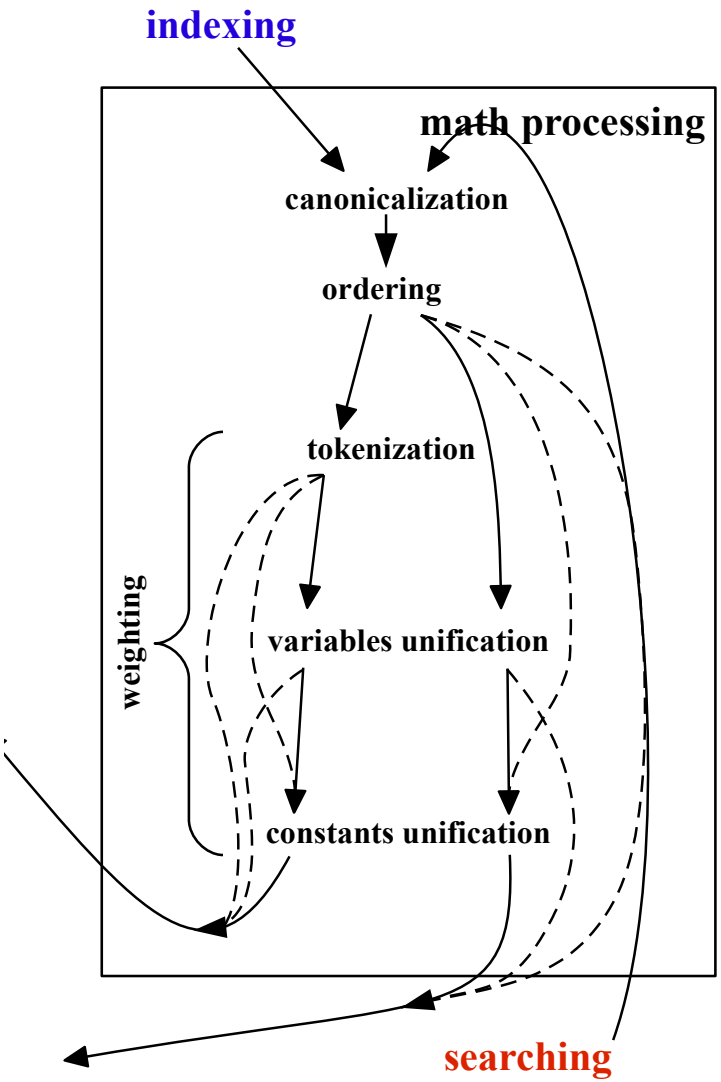


indexing

$x^y + y^3$

searching

Example



indexing

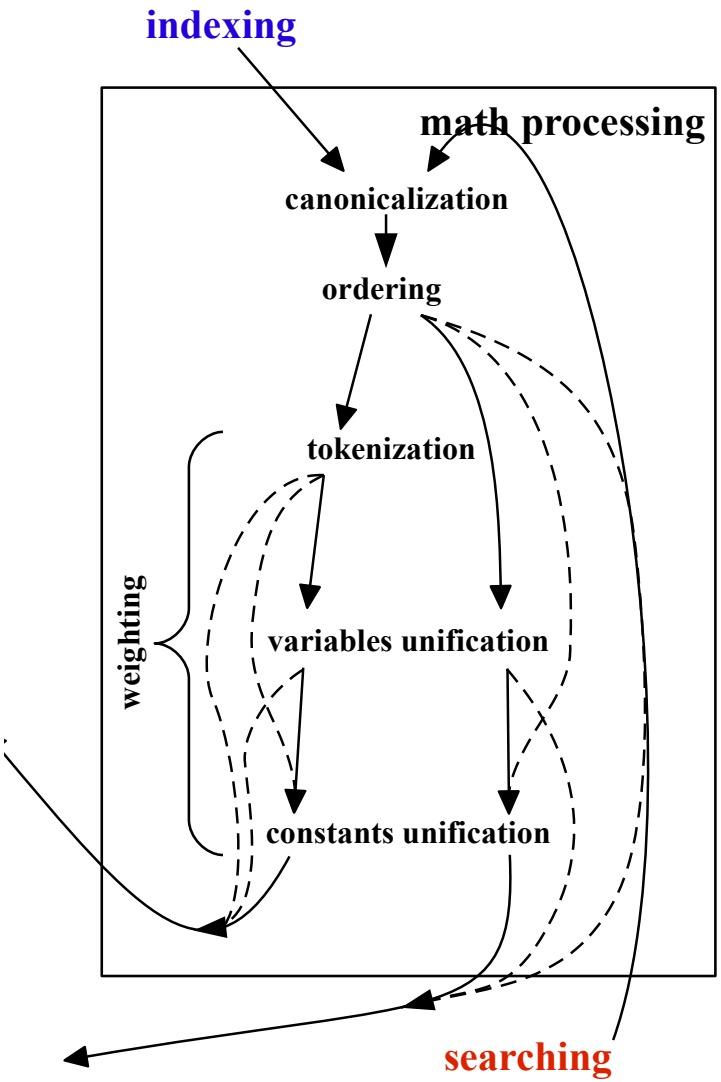
$x^y + y^3$

↓

$x^y + y^3$

searching

Example



indexing

$x^y + y^3$

↓

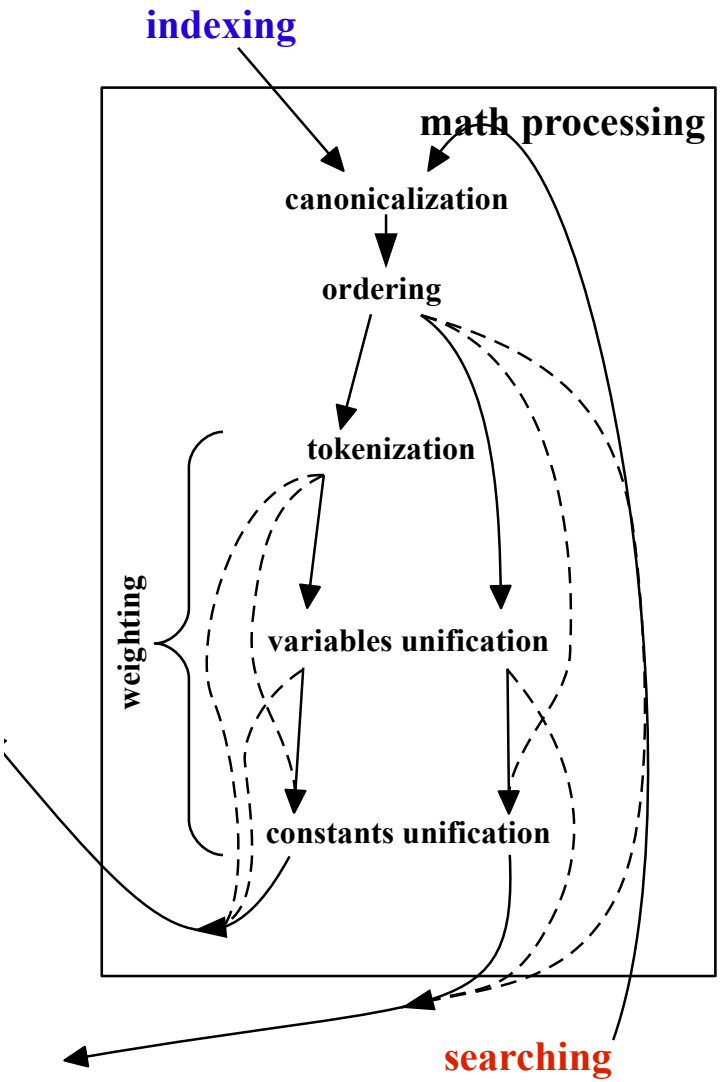
$x^y + y^3$

↓

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

searching

Example



indexing

x^y+y^3

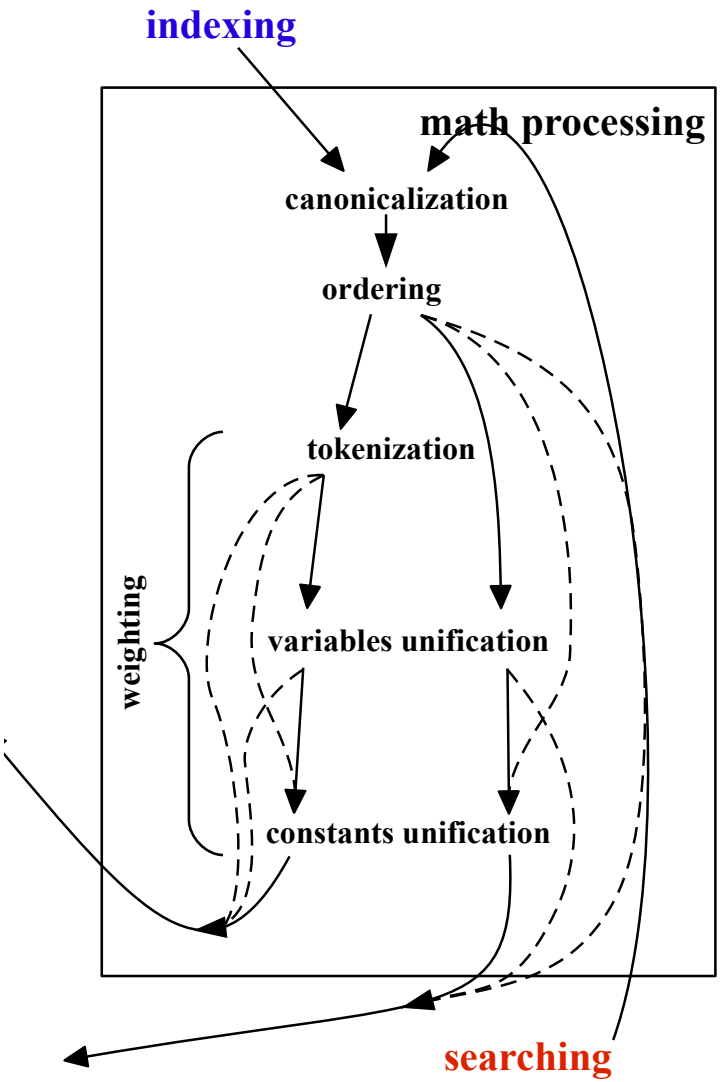
x^y+y^3

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

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

searching

Example



indexing

$$x^y+y^3$$

$$\downarrow$$

$$x^y+y^3$$

$$\downarrow$$

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

$$\downarrow$$

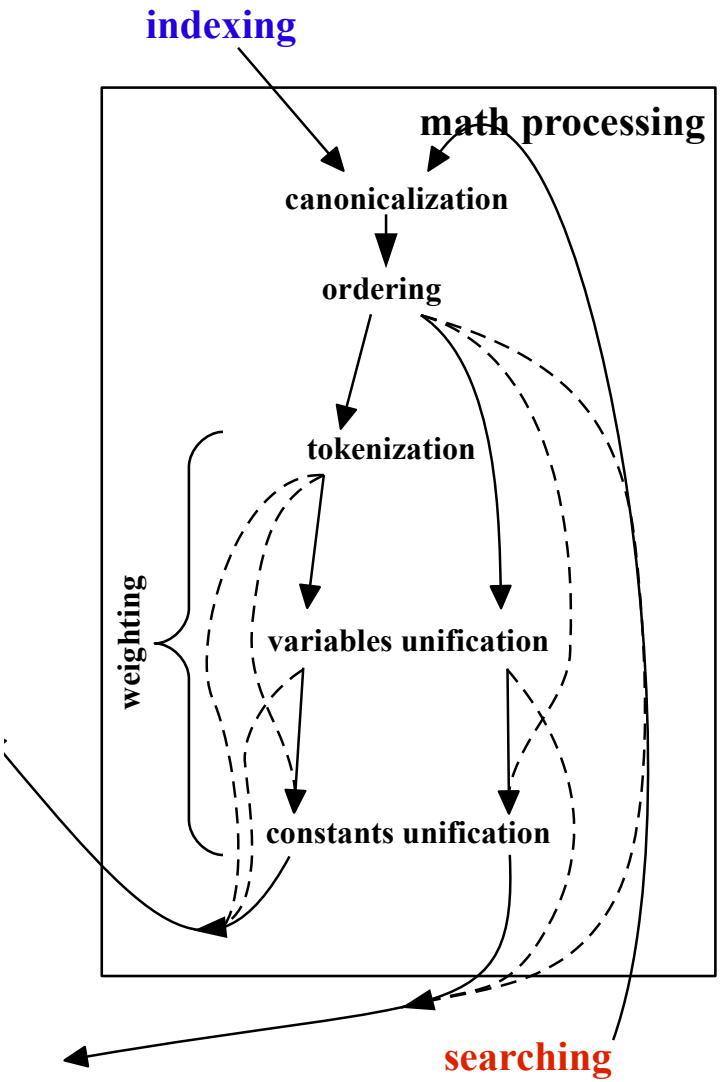
$$x^y+y^3\ ,\ x^y\ ,\ y^3\ ,\ x\ ,\ y\ ,\ 3,+\ ,\ id_1^{id_2}+id_2^3,\ id_1^{id_2},\ id_1^3$$

$$\downarrow$$

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

searching

Example



indexing

x^y+y^3

x^y+y^3

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

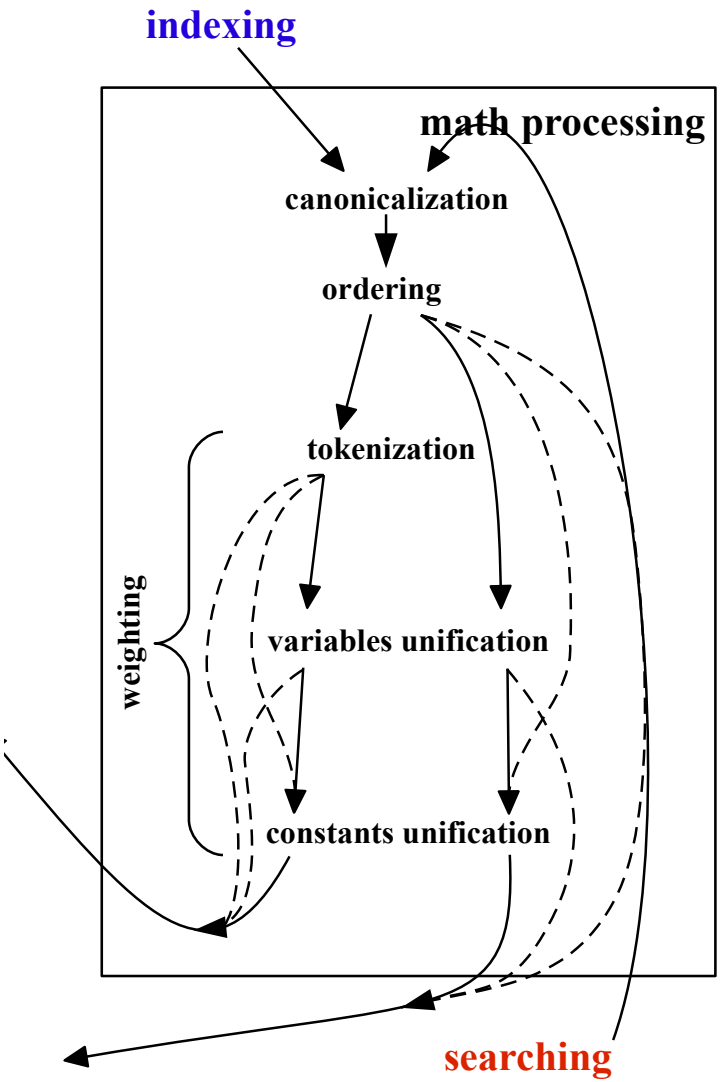
$x^y+y^3\ ,\ x^y\ ,\ y^3\ ,\ x\ ,\ y\ ,\ 3,+\ ,\ id_1^{id_2}+id_2^3,\ id_1^{id_2},\ id_1^3$

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

searching

x^y+y^2

Example



indexing

x^y+y^3

x^y+y^3

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

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

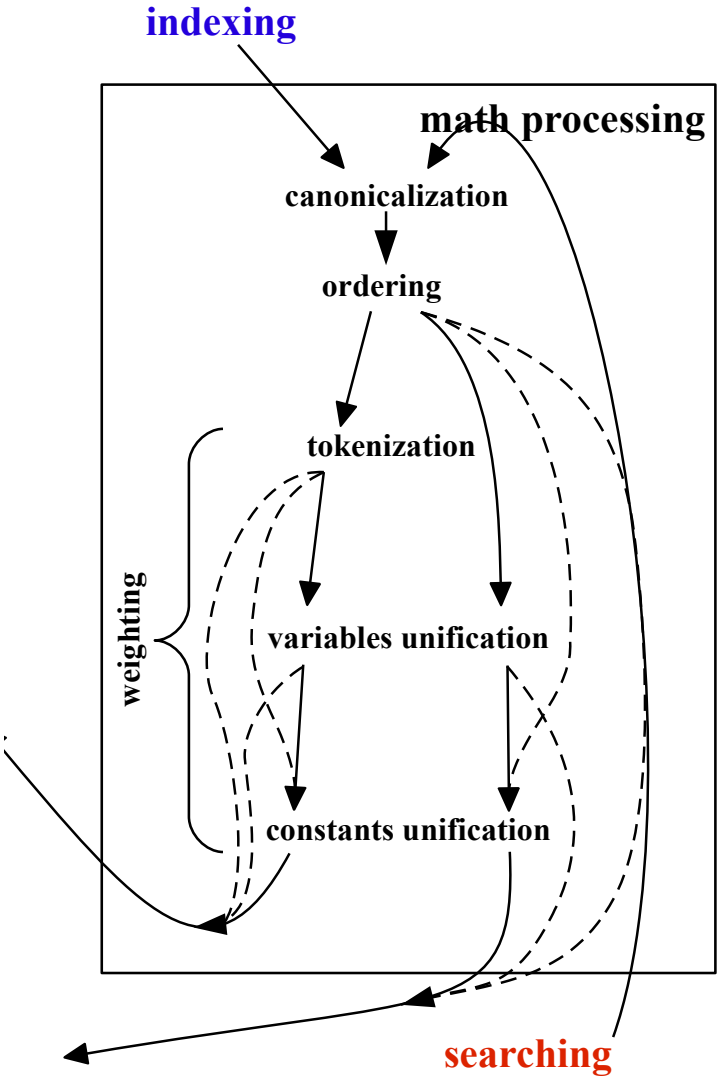
$x^y+y^3\ ,\ x^y\ ,\ y^3\ ,\ x\ ,\ y\ ,\ 3,+\ ,\ id_1^{id_2}+id_2^3,$
$id_1^{id_2},\ id_1^3,\ x^y+y^{const}\ ,\ y^{const}\ ,\ id_1^{id_2}+id_2^{const}\ ,\ id_1^{const}$

searching

x^y+y^2

x^y+y^2

Example



indexing

x^y+y^3

x^y+y^3

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

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

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

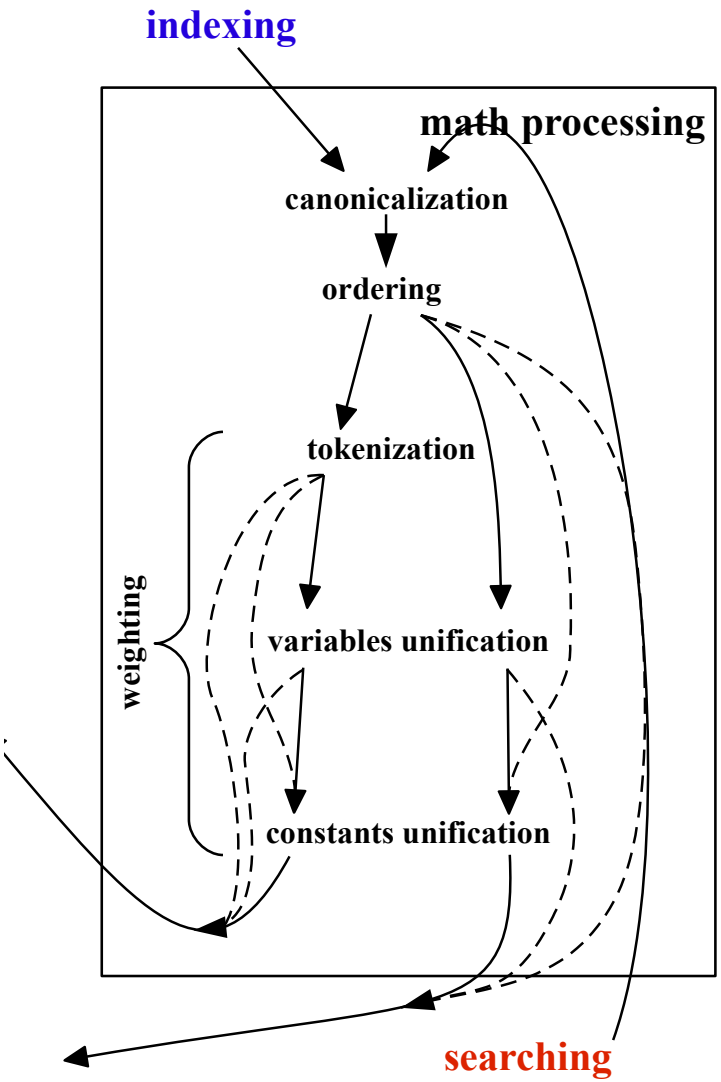
searching

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

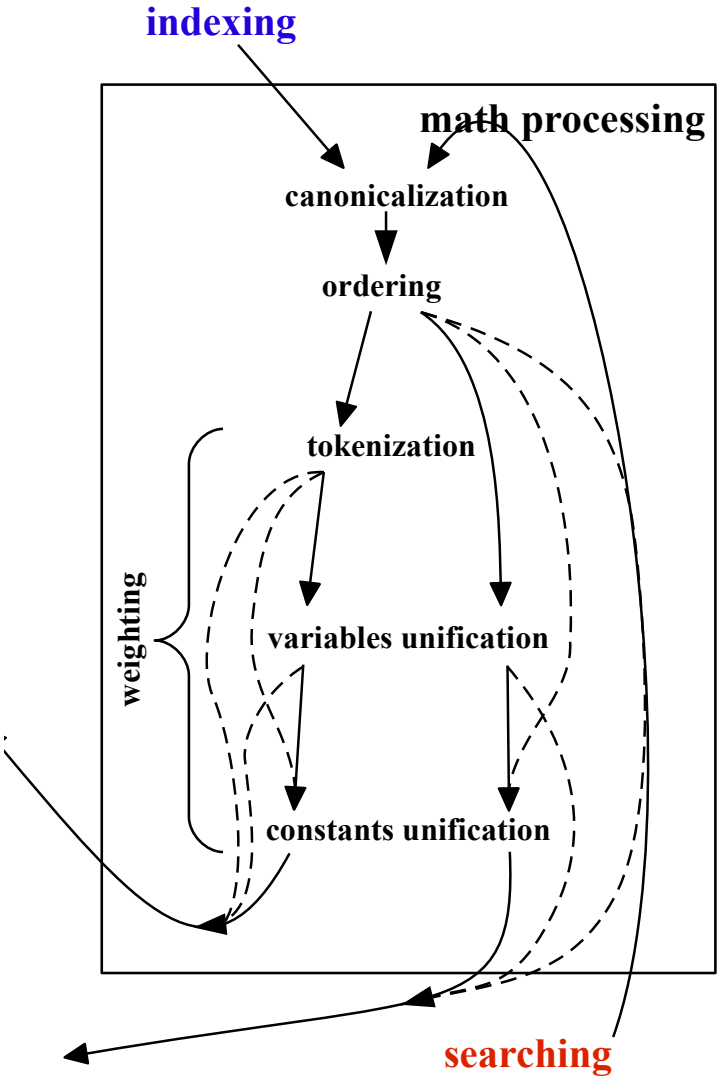
x^y+y^2

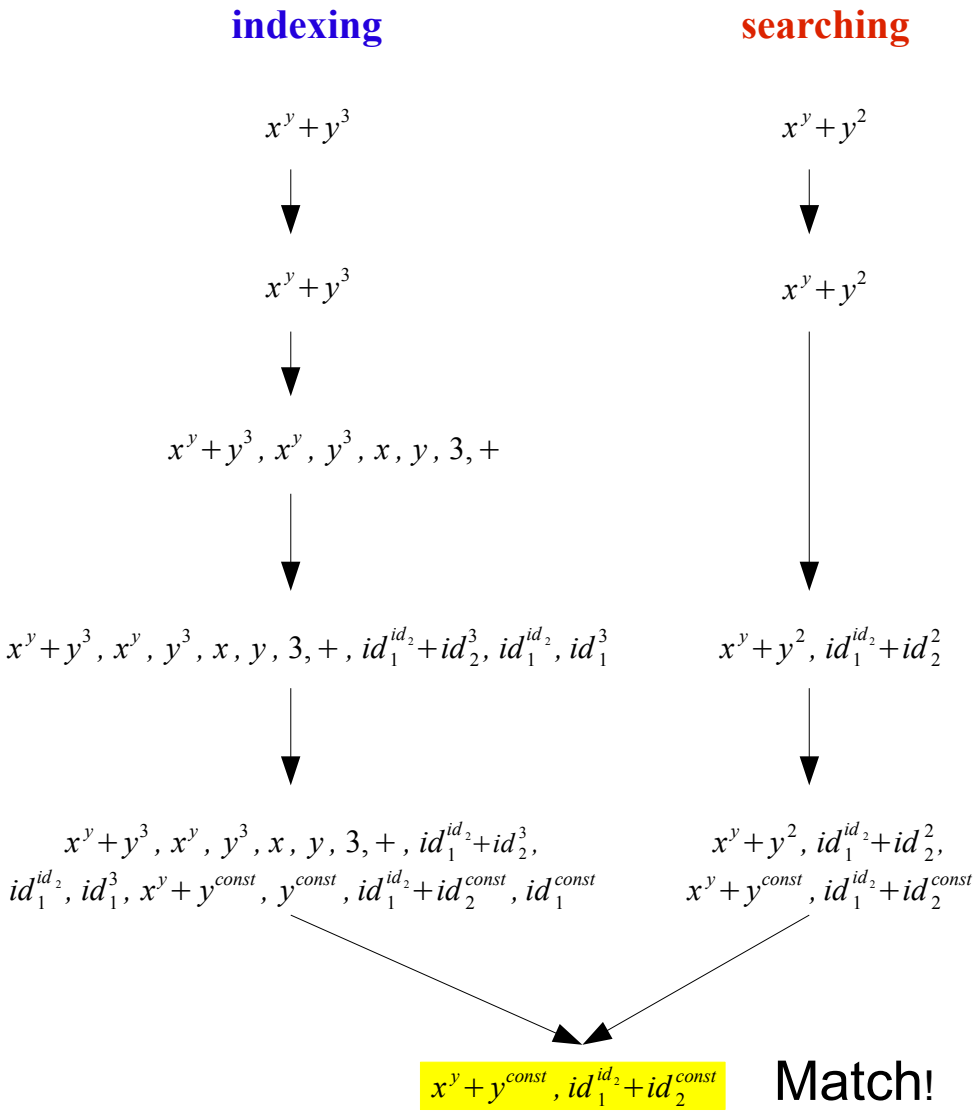
x^y+y^2

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

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

Example


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


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

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

Formula Processing Example

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

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

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

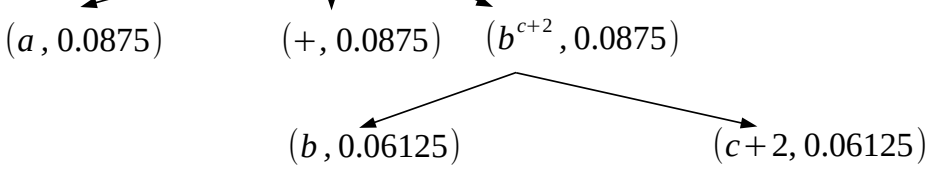
$$0.0875 = 0.125 \cdot 0.7^{(l)}$$

Formula Processing Example

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

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

tokenization:



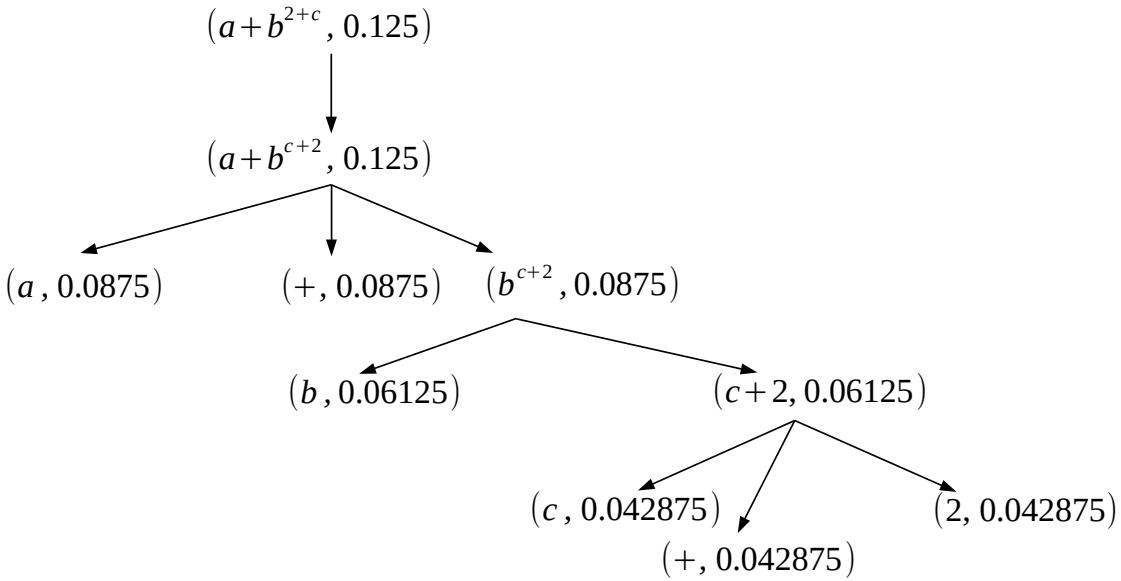
$$0.06125 = 0.0875 \cdot 0.7(l)$$

Formula Processing Example

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

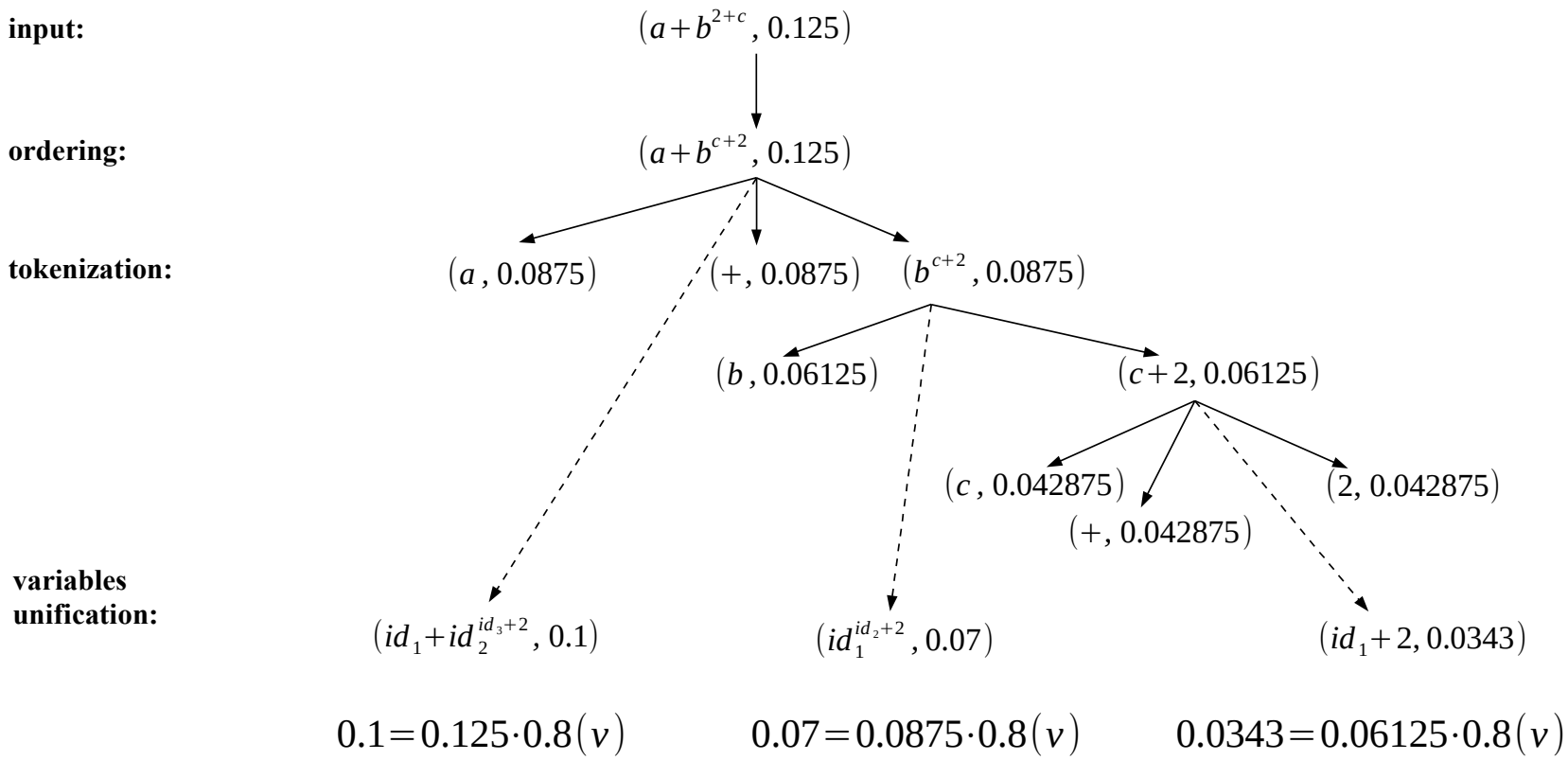
ordering: $(a+b^{c+2}, 0.125)$

tokenization:

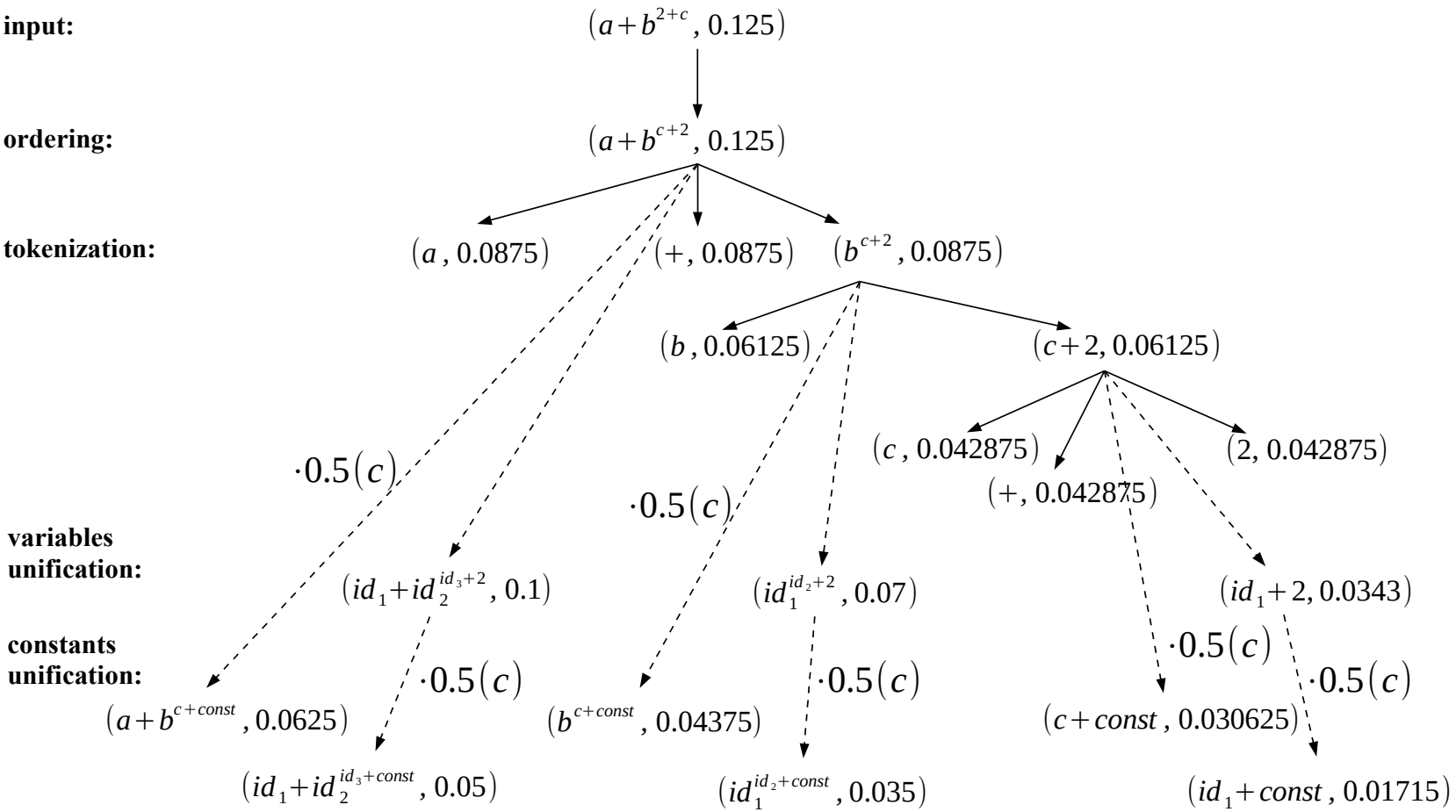


$$0.042875 = 0.06125 \cdot 0.7(l)$$

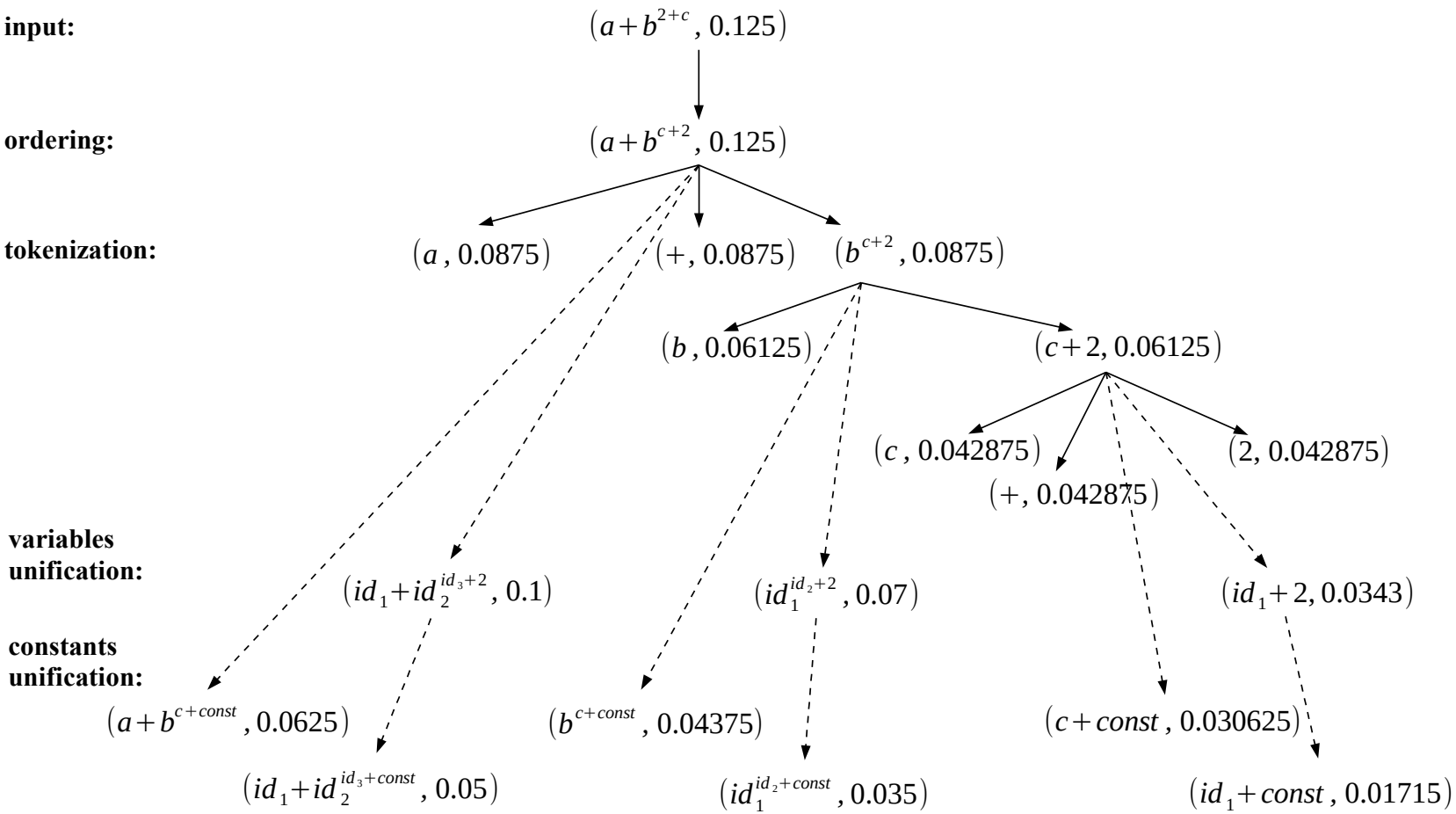
Formula Processing Example



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
  - Uncompressed size 124 GB, compressed 15 GB
  - 158 million input formulae, 2.9 billion expressions indexed

MIR

- MIR sandbox
  - 10 000 documents
  - Uncompressed size 1.75 GB, compressed 367 MB
  - 1.6 million input formulae, 22 million expressions indexed

MIR

- MIR harvests
  - 119 documents
  - Uncompressed size 1.24 GB, compressed 201 MB
  - 1.19 million input formulae, 21 million expressions indexed

WebMIaS


Examples About Help Contact
WEBMIaS
MATH INDEXER AND SEARCHER
x^2+y^2 exponential distribution
Search in: MREC 2011.4.439
Search
Total hits: 15970, showing 1- 30. Searching time: 112 ms
Estimating copula measure using ranks and subsampling: a simulation study
For the dependence 3, we will test use the Komogorov-Smirnov test to know whether x^2+y^2 is exponentially distributed (true if ...
score = 0.04348715
arxiv.org/abs/0709.3860 - cached XHTML
Real-time TPC Analysis with the ALICE High-Level Trigger
... $\sqrt{x^2+y^2}$ ...
score = 0.04333227
arxiv.org/abs/physics/0403063 - cached XHTML
Pairing symmetry and long range pair potential in a weak coupling theory of ...
... does not mix with usual $s_{x^2+y^2}$ symmetry gap in an anisotropic band structure.
score = 0.03675753
arxiv.org/abs/cond-mat/9906142 - cached XHTML

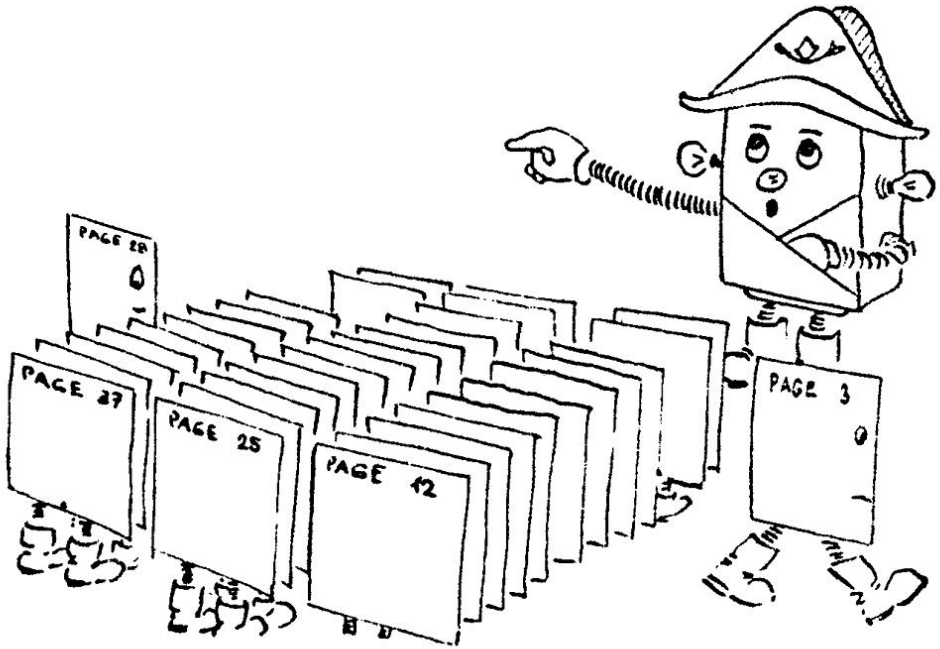
WebMIaS

- Demo web interface: http://nlp.fi.muni.cz/projekty/eudml/mias
- MIR development demo: http://aura.fi.muni.cz:8085/webmias-mir/
  - MathML/TEX input (Tralics for conversion to MathML)
  - Matched document snippet generation
  - MathJax for nicer math rendering and better portability

Conclusion

- Project pages – http://nlp.fi.muni.cz/projekty/eudml/mias
- Future work
  - Further canonicalization and unification
  - Optimization
  - Mathematical equivalence computation via symbolic algebra system?
  - Suggestions welcome!

Questions?




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 [5], 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 [5], 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://www.fi.muni.cz/ sojka/dml-2011-program.html>.