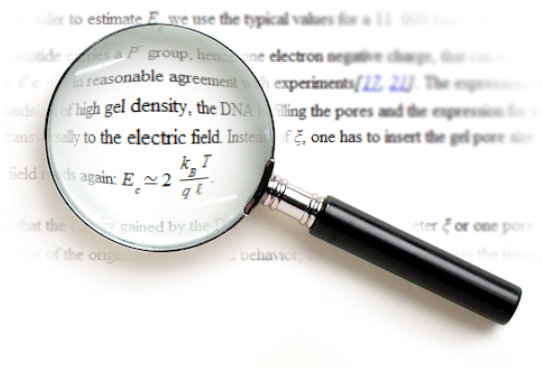
Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task

Martin Líška, Petr Sojka, Michal Růžička

Masaryk University, Faculty of Informatics, Brno, Czech Republic

June 21st, 2013



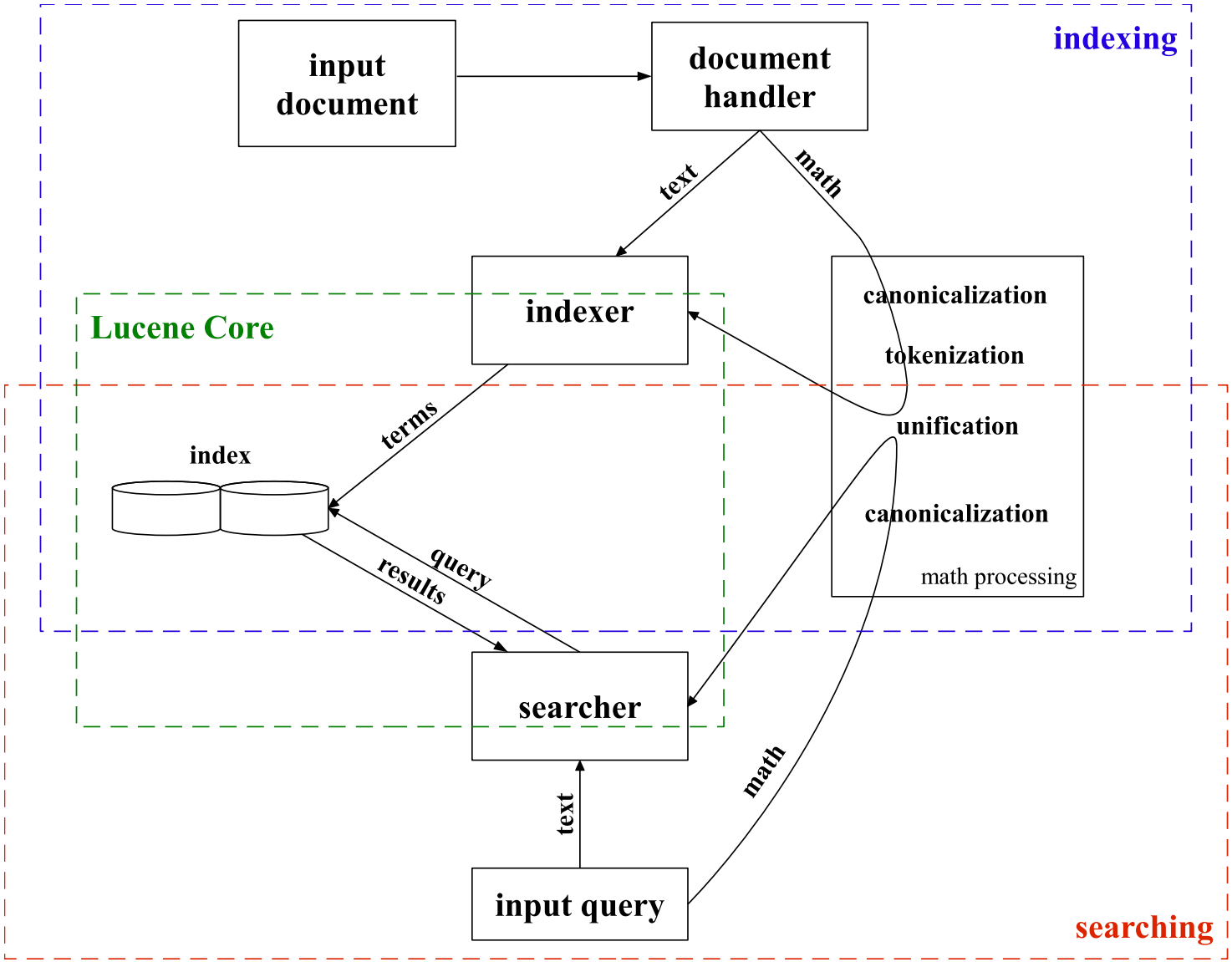
Motivation

- Digital Mathematics Libraries – EuDML, DML-CZ, …
- Knowledge in mathematical expressions
- Example
- MIaS evaluation
  - How well does it work?

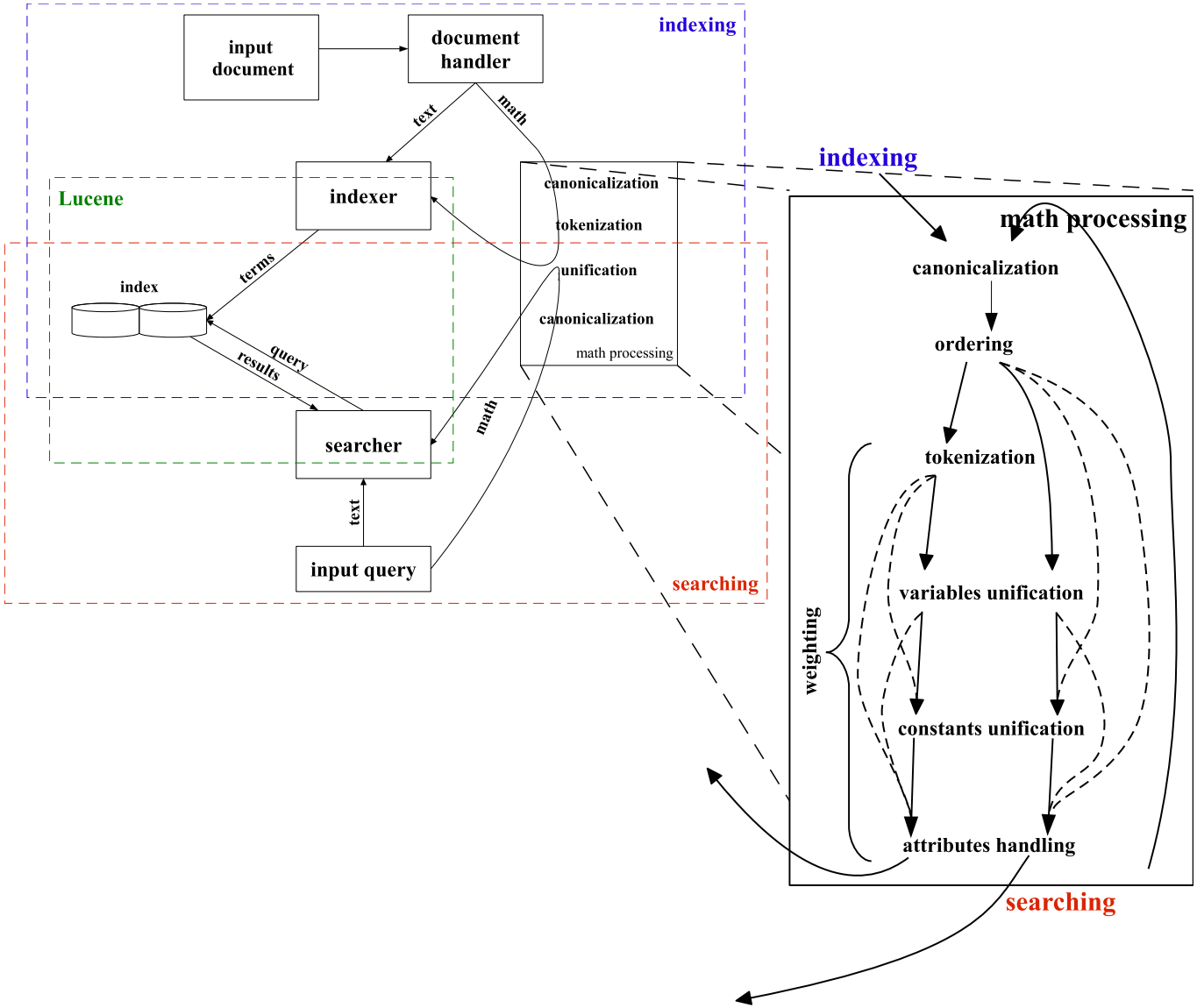
Math Indexer and Searcher (MIaS)

- Similarities of math expression trees (*both* structural (PMath) and semantic (CMath))
- Full-text based, document oriented, math-aware search engine
- Mixed text-math indexing and searching
- Query language: MathML, LaTeX
- Results ordering based on similarity level
- Snippets with hit highlighting
- On-the-fly query formula rendering

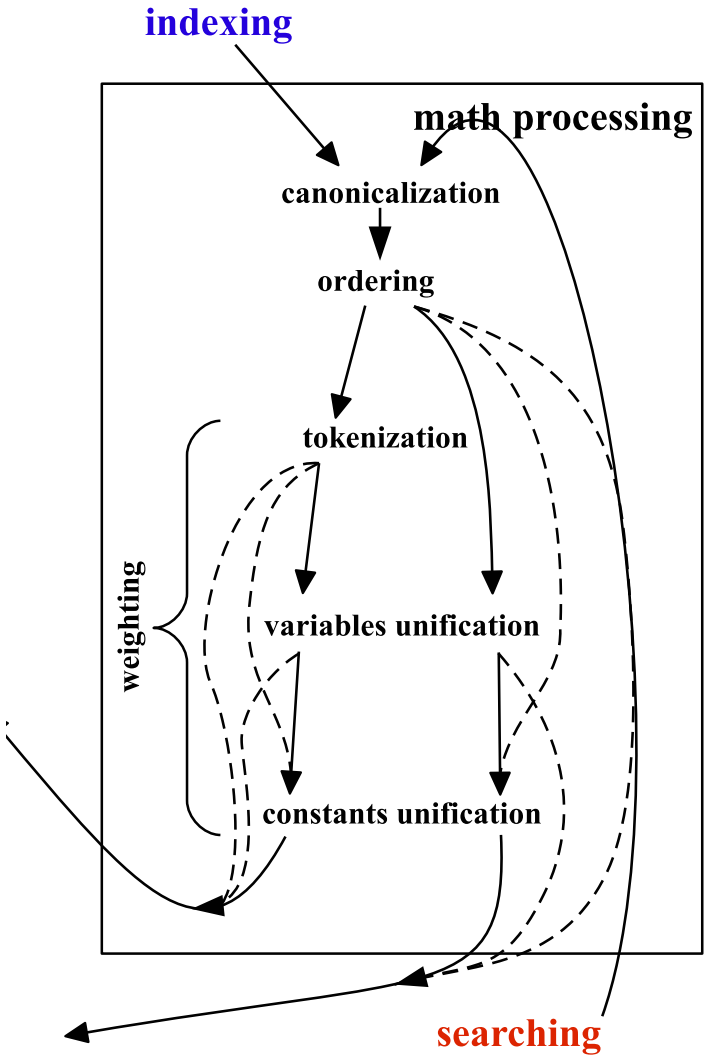
Design

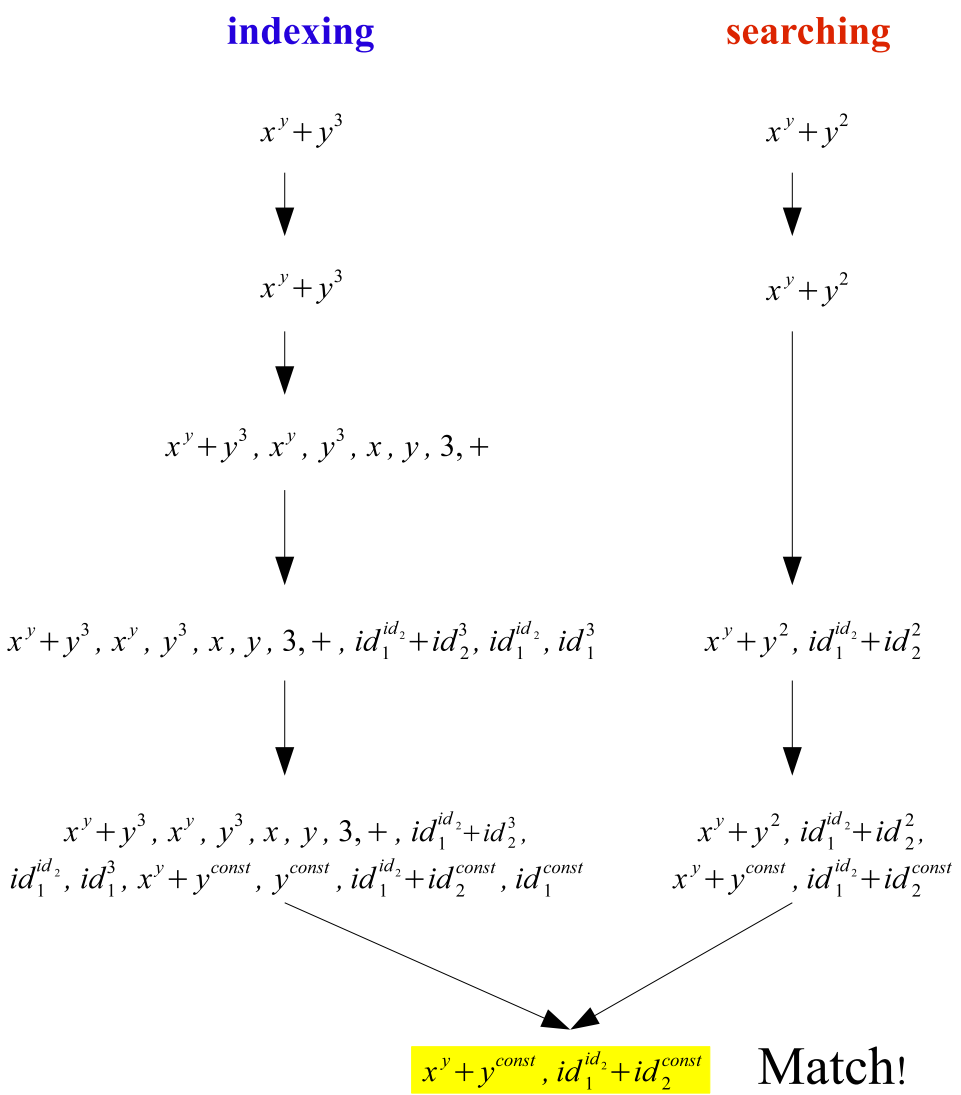

indexing
input document
document handler
text
math
indexer
Lucene Core
canonicalization
tokenization
unification
canonicalization
math processing
terms
index
query
results
searcher
text
math
input query
searching

Design II


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

Example


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


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

MIaS – Math Task

- Formula search and Full-text search
  - 4 runs submitted – differ in query language
    - PMath – Run #1
    - CMath – Run #2
    - PCMath – Run #3
    - TEX – Run #4
- Open Information Retrieval
  - 1 run submitted – TEX+text mixed queries

Querying

- Automatic querying scripts, generated:
  - trec_eval format
  - XML
  - Clickable HTML for investigation and fine tuning

Querying


Results for 'TeX' run with query ID 'NTCIR10-FS-21' in index '1' (XML response)
Results for 'TeX' run with query ID 'NTCIR10-FS-22' in index '1' (XML response)
Results for 'PMath' run with query ID 'NTCIR10-FT-1' in index '0' (XML response)
WEBMIaS
MATH INDEXER AND SEARCHER
Your query:
Converted and canonicalized query:
This XML file does not appear to have any style information associated with it. The document tree is shown below.

Results

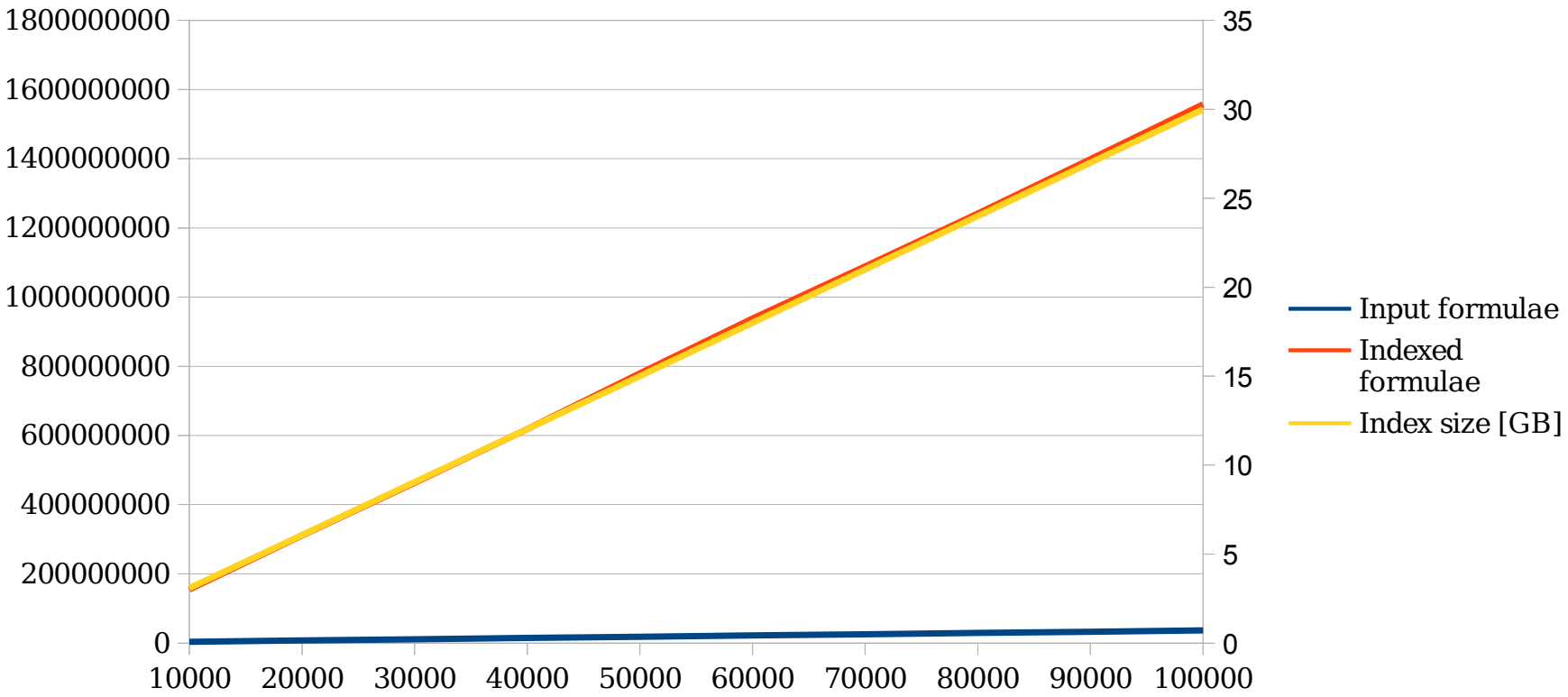
Table 1: Result metrics for submitted runs in Formula Search with Relevance Level ≥ 3 (Relevant)

Metric	Run 1	Run 2	Run 4
P-10 avg	0.105	0.191	**0.219**
P-5 avg	0.133	0.229	**0.276**
MAP avg	0.060	0.112	**0.127**
Precision	0.109 (64/589)	**0.185** (92/496)	0.123 (96/778)

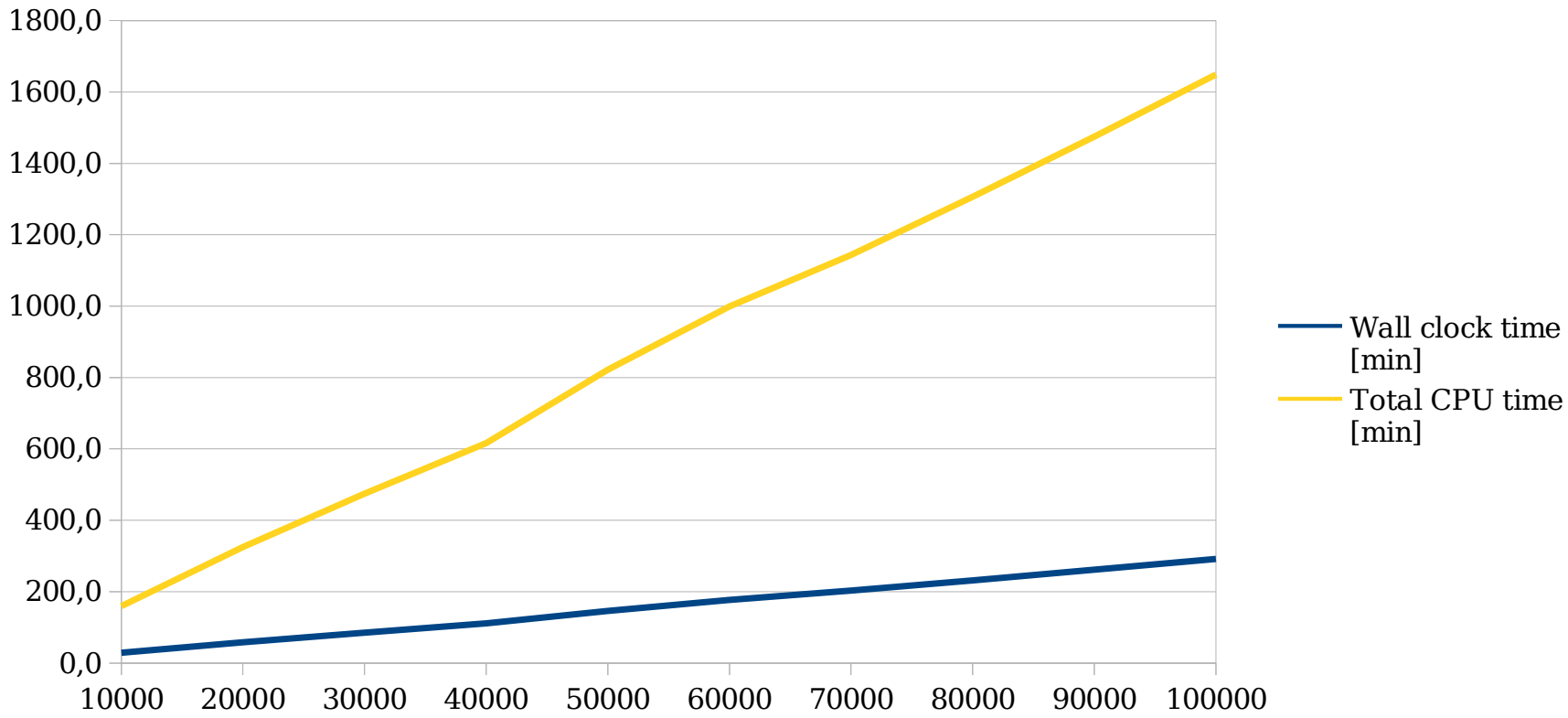
Table 2: Result metrics for submitted runs in Formula Search with Relevance Level ≥ 1 (Partially Relevant)

Metric	Run 1	Run 2	Run 4
P-10 avg	0.143	0.214	**0.267**
P-5 avg	0.181	0.267	**0.343**
MAP avg	0.066	0.081	**0.100**
Precision	0.148 (87/589)	**0.232** (115/496)	0.161 (125/778)

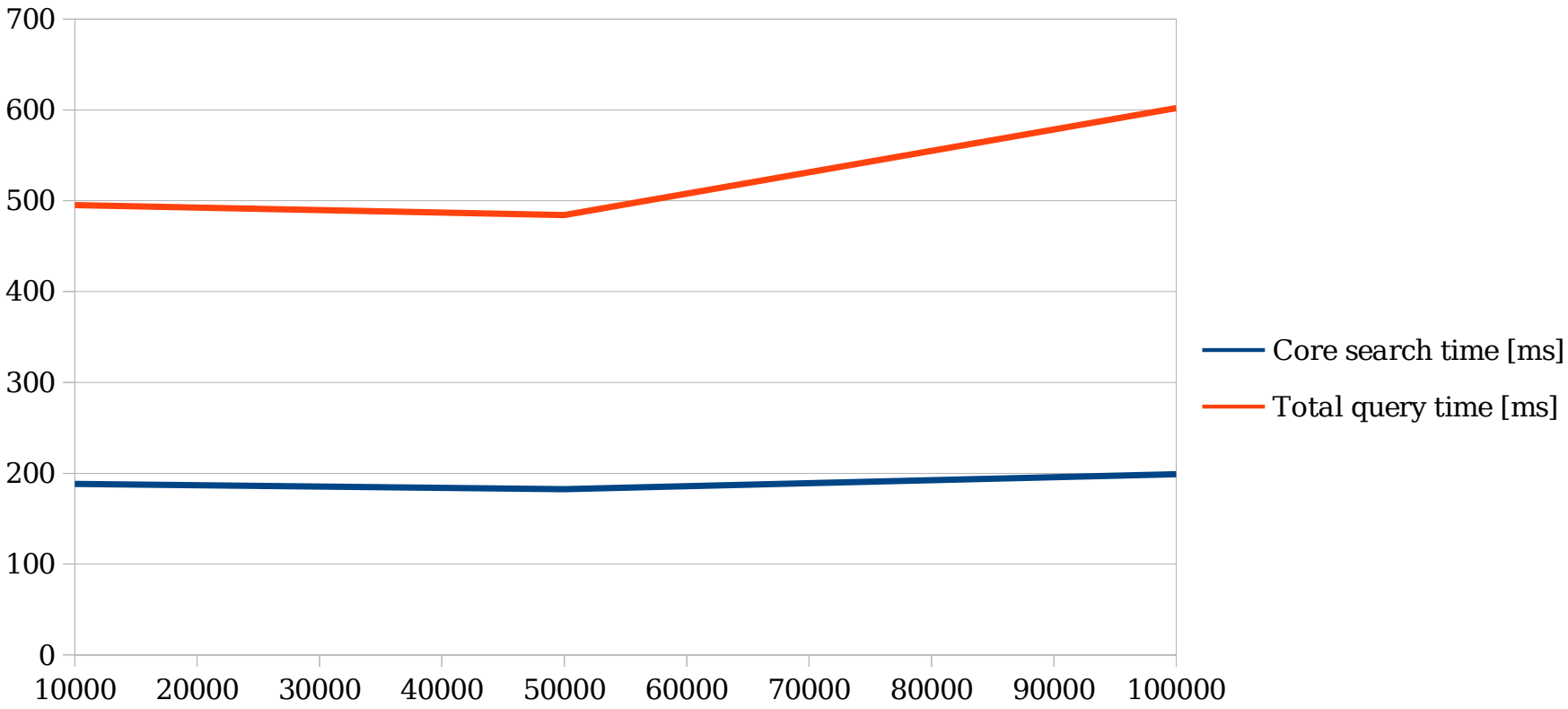
Efficiency


1800000000
1600000000
1400000000
1200000000
1000000000
800000000
600000000
400000000
200000000
0
35
30
25
20
15
10
5
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
Input formulae
Indexed formulae
Index size [GB]

Efficiency


1800,0
1600,0
1400,0
1200,0
1000,0
800,0
600,0
400,0
200,0
0,0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
Wall clock time [min]
Total CPU time [min]

Efficiency


700
600
500
400
300
200
100
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
Core search time [ms]
Total query time [ms]

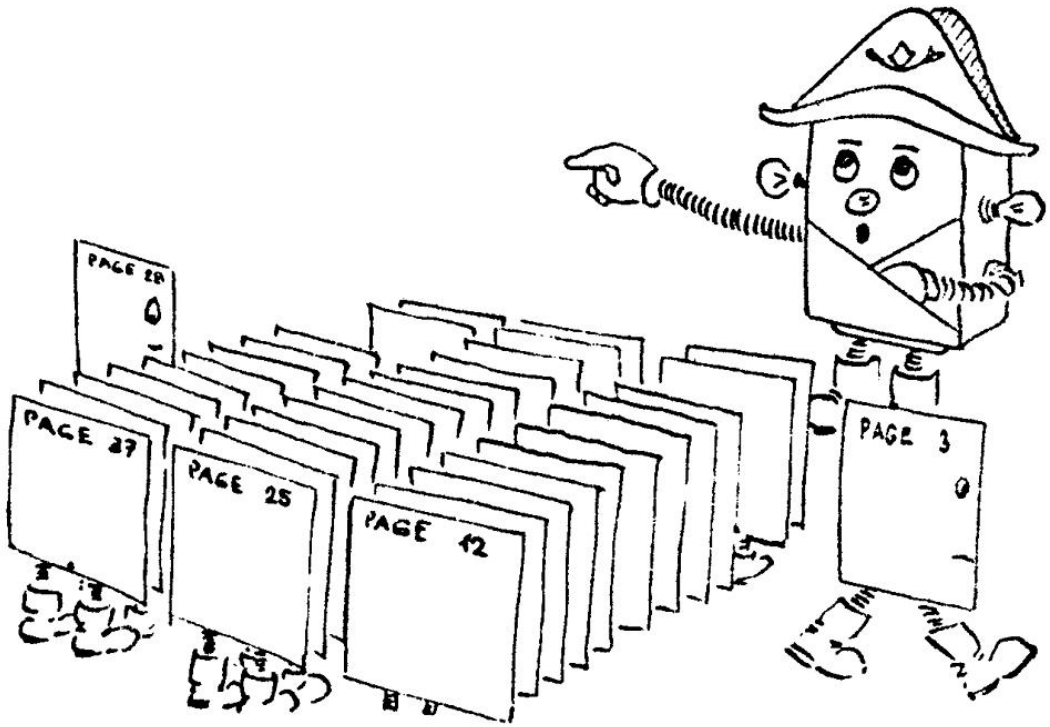
Conclusions

- Evaluation is driving the development
- Our participation to NTCIR shown useful
- TEX queries (both PMath and CMath used) most precise
- Deeper investigation of the results

Future work

- Own evaluation framework (evaluation driven development)
- Canonicalization
- Experiment with weight factors and formulae relevance weighting
- Content (aka semantics) focused search
- Analysis from running system (EuDML) and its logs

Questions?




Líška, Martin and Petr Sojka and Michal Růžička. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math T ask. In Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Math Pilot Task


D. Formánek, M. Líška, M. Růžička, and P. Sojka. Normalization of digital mathematics library content. In J. Davenport, J. Jeuring, C. Lange, and P. Libbrecht, editors, 24th OpenMath Workshop, 7th Workshop on Mathematical User Interfaces (MathUI), and Intelligent Computer Mathematics Work in Progress, number 921 in CEUR Workshop Proceedings, pages 91–103, Aachen, 2012.


Sojka, Petr and Martin Líška. The Art of Mathematics Retrieval. In Matthew R. B. Hardy , Frank Wm. T ompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View , CA, USA: ACM, 2011. p. 57–60, 4 pp. ISBN 978-1-4503-0863-2. <http://dx.doi.org/10.1145/2034691.2034703>


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>


Sojka, P. (ed.): Towards a Digital Mathematics Library. Masaryk University, Paris, France (Jul 2010), <http://www.fi.muni.cz/ sojka/dml-2010-program.html>


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>.