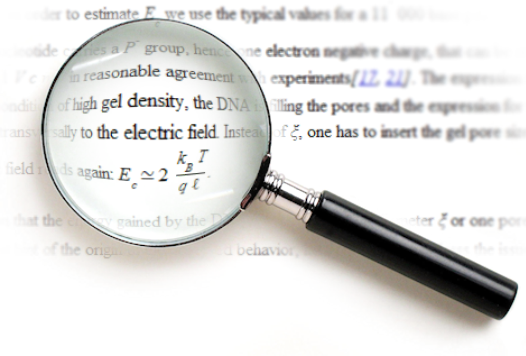
Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy

Michal Růžička, Petr Sojka, Martin Líška

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

https://mir.fi.muni.cz/



Illustrations by Jiří Franek.

Outline

Outline

NTCIR-10 Math Task

- The first (pilot) year of the math task event last year (i.e. 2013).
- 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.

NTCIR-10 Math Task 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)

NTCIR-11 Math-2 Task

- Only one type of queries.
  - 50 queries, each
    - 1–4 formulae,
    - 1–4 keyphrases.
- Wikipedia task in addition to the Main task.

NTCIR-11 Math-2 Main Task Results

Table: Results of submitted runs with Relevance Level ≥ 3 (Relevant). Main task team rank is in [ ] for our best runs (in bold).

	PMath	CMath	PCMath	TEX
MAP avg	0.3073	**0.3630 [1]**	0.3594	0.3357
P@10 avg	0.3040	**0.3520 [1]**	0.3480	0.3380
P@5 avg	0.5120	**0.5680 [1]**	0.5560	0.5400

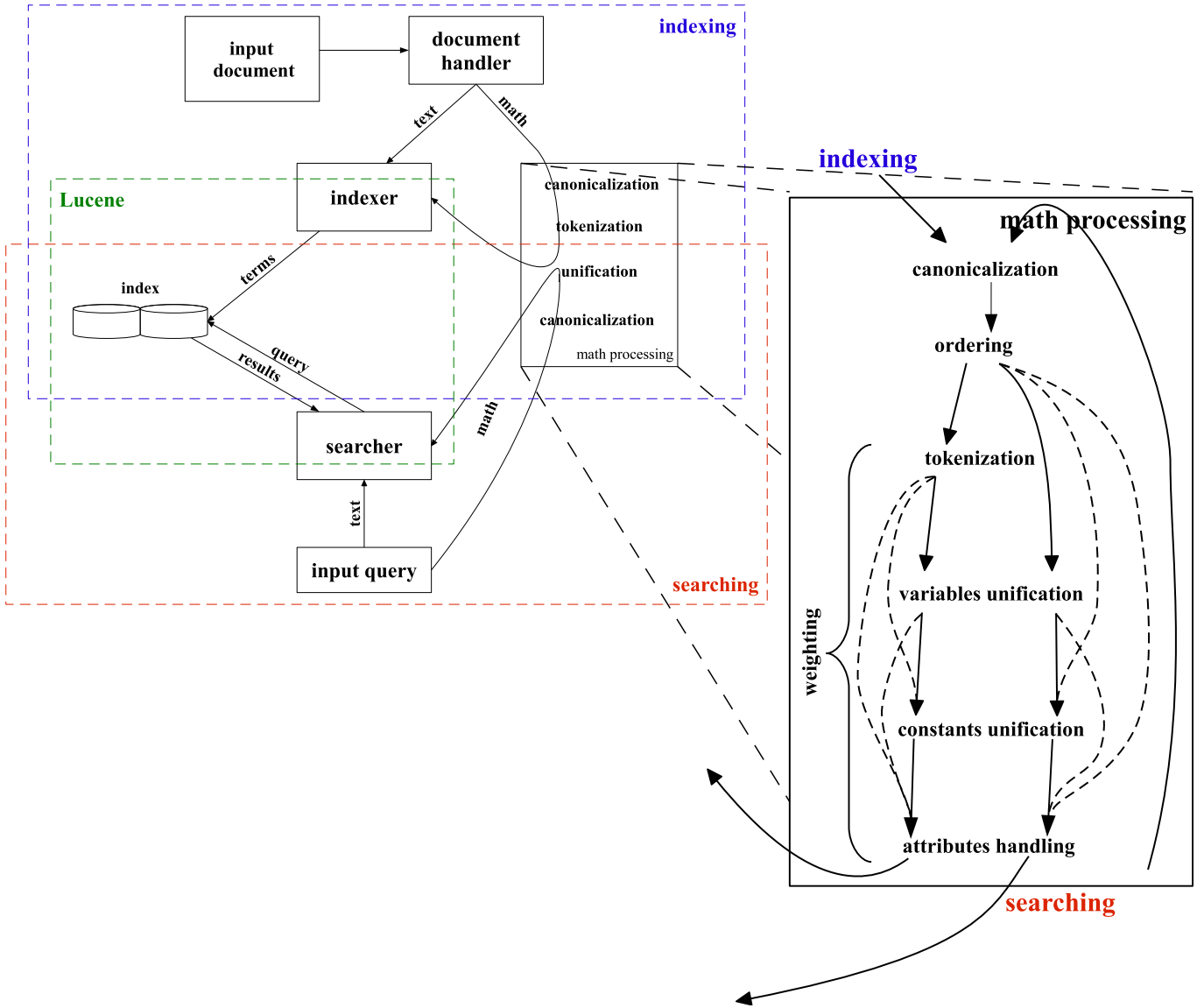
Table: Results of submitted runs with Relevance Level ≥ 1 (Partially Relevant). Number in [ ] is team rank of all runs.

	PMath	CMath	PCMath	TEX
MAP avg	0.2557	**0.2807 [2]**	0.2799	0.2747
P@10 avg	0.5020	0.5440	**0.5520 [1]**	0.5400
P@5 avg	0.8440	**0.8720 [2]**	0.8640	0.8480

NTCIR-11 Math-2 Wikipedia Task Results

- Topics with results:
  - 75 out of 100 (CMath run)
- Average position:
  - 64 correct results in top 100
  - 58 correct results in top 20
  - 56 correct results in top 10
  - 53 correct results in top 5
  - 52 correct results in top 4
  - 50 correct results in top 3
  - 48 correct results in top 2
  - 46 correct results in top 1

NTCIR-11 Math-2 Main Task Approach



NTCIR-11 Math-2 Main Task Approach: News

- Query expansion & strip-merging of subresults.
  - Query expansion.

query 1 (the original query):	f_1	f_2	k_1	k_2	k_3
query 2:	f_1	f_2	k_1	k_2	
query 3:	f_1	f_2	k_1		
query 4:	f_1	f_2			
query 5:	f_1		k_1	k_2	k_3
query 6:			k_1	k_2	k_3

NTCIR-11 Math-2 Main Task Approach: News

- Strip-merging of subresults.
  - Example on three subqueries (the original one and two derived subqueries).

Results of the original query:

1: $r1_{\text{original}}$
2: $r2_{\text{original}}$
3: $r3_{\text{original}}$
4: $r4_{\text{original}}$
5: $r5_{\text{original}}$
6: $r6_{\text{original}}$
7: $r7_{\text{original}}$
8: $r8_{\text{original}}$
9: $r9_{\text{original}}$
10: $r10_{\text{original}}$
11: $r11_{\text{original}}$

Results of the subquery 1:

1: $r1_{\text{subquery 1}}$
2: $r2_{\text{subquery 1}}$
3: $r3_{\text{subquery 1}}$
4: $r4_{\text{subquery 1}}$
5: $r5_{\text{subquery 1}}$

Results of the subquery 2:

1: $r1_{\text{subquery 2}}$
2: $r2_{\text{subquery 2}}$
3: $r3_{\text{subquery 2}}$
4: $r4_{\text{subquery 2}}$
5: $r5_{\text{subquery 2}}$

The final result list:

1: $r1_{\text{original}}$
2: $r2_{\text{original}}$
3: $r3_{\text{original}}$
4: $r1_{\text{subquery 1}}$
5: $r2_{\text{subquery 1}}$
6: $r1_{\text{subquery 2}}$
7: $r4_{\text{original}}$
8: $r5_{\text{original}}$
9: $r6_{\text{original}}$
10: $r3_{\text{subquery 1}}$
11: $r4_{\text{subquery 1}}$
12: $r2_{\text{subquery 2}}$
13: $r7_{\text{original}}$
14: $r8_{\text{original}}$
15: $r9_{\text{original}}$
16: $r5_{\text{subquery 1}}$
No more results from subquery 1.
17: $r3_{\text{subquery 2}}$
18: $r10_{\text{original}}$
19: $r11_{\text{original}}$
No more results from the original query.
20: $r4_{\text{subquery 2}}$
21: $r5_{\text{subquery 2}}$
No more results from subquery 2.
22: $r1_{\text{random}}$
23: $r2_{\text{random}}$
...
1000: $r979_{\text{random}}$

Query Expansion Results' Insight

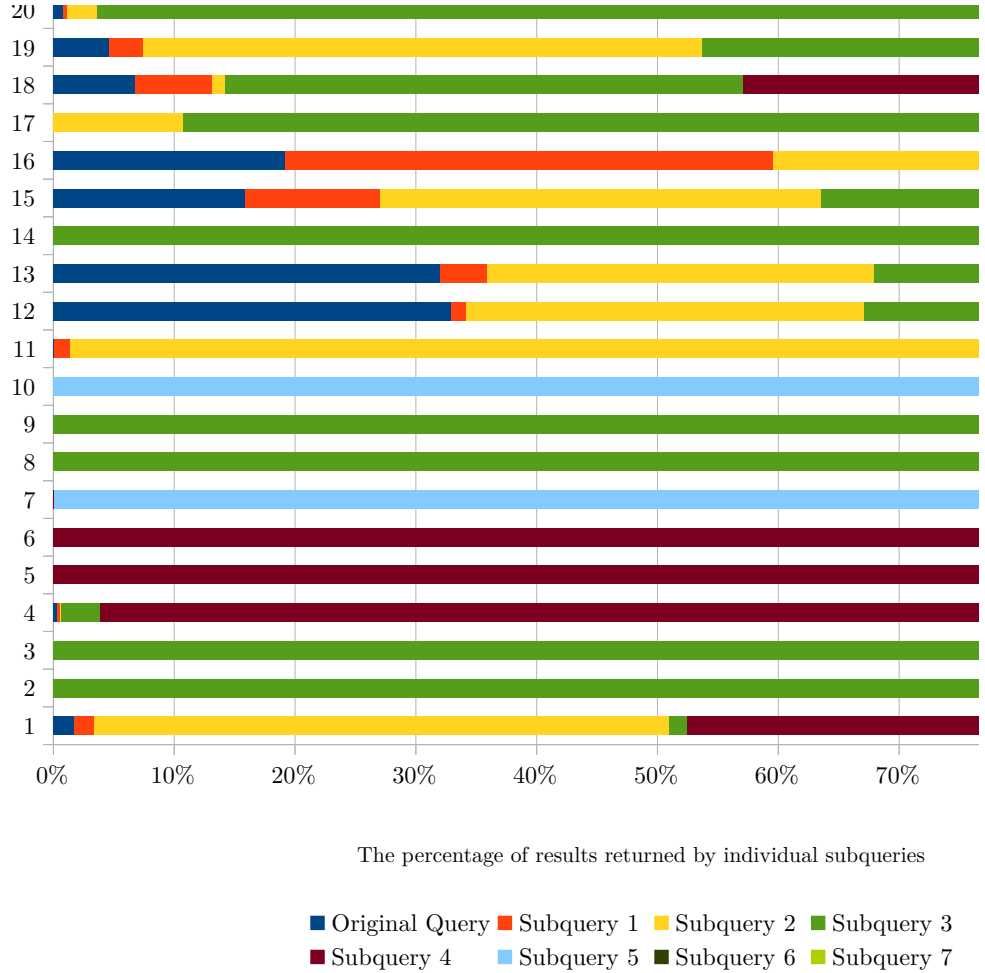


Figure: Relative number of results found using different subqueries for every query in CMath run

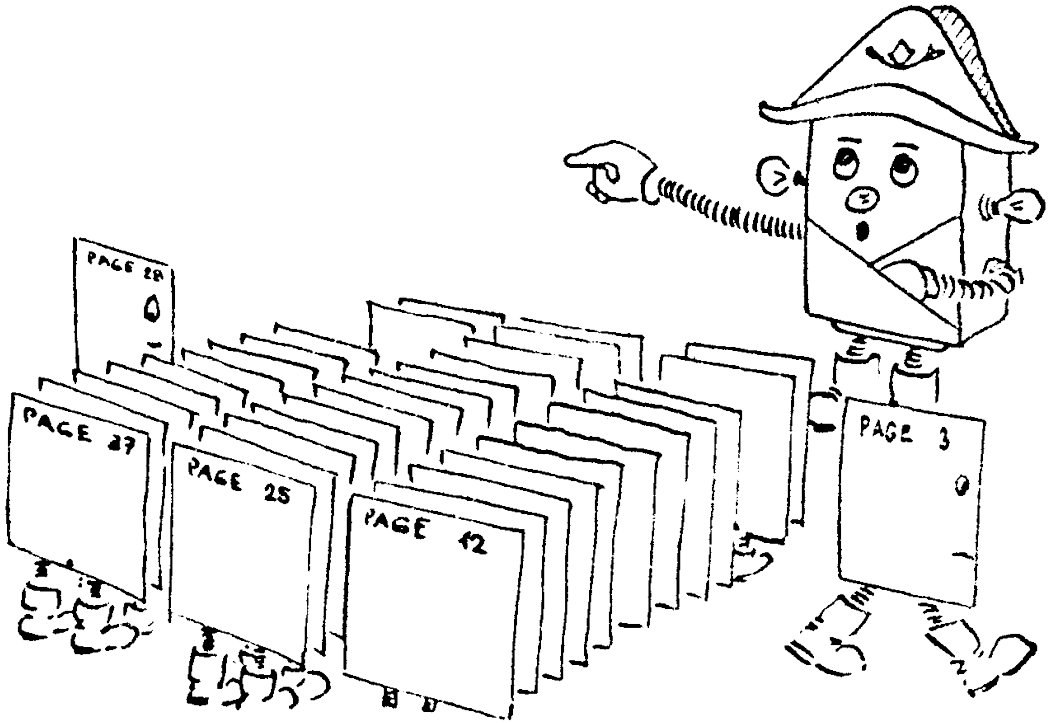
NTCIR-11 Math-2 Wikipedia Task Content Topics

- Completely the same fully automatic system used for the main NTCIR Math Task and Wikipedia subtask.
  - Only different data.
  - No tuning or modifications for the Wikipedia task.
- Input Content MathML was transformed to the format of the main NTCIR math task.
  - Manually added Presentation MathML and TeX representation of the data.
  - Performed all the four runs (CMath, PMath, PCMath, TeX) similarly to the main task.
- No query expansion & strip-merging possible as queries consist of a single formula only.

Summary

- Our results significanlty *improved* since the last year.
- Query expansion & strip-merging of subresults helps *a lot*.
- Better unification *definitely* needed.
- Wikipedia task *very* useful.

Questions?




Illustrations by Jiří Franek.


SOJKA, Petr and Martin LÍŠKA. The Art of Mathematics Retrieval. In Matthew R. B. Hardy, Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View, CA, USA: ACM, 2011. p. 57–60. ISBN 978-1-4503-0863-2. doi:10.1145/2034691.2034703.


LÍŠKA, Martin, Petr SOJKA and Michal RŮŽIČKA. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task. In Noriko Kando, Kazuaki Kishida. Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan, 2013. s. 686-691, 6 s. ISBN 978-4-86049-062-1.


LÍŠKA, Martin, Petr SOJKA, Michal RŮŽIČKA and Peter MRAVEC. Web Interface and Collection for Mathematical Retrieval : WebMIaS and MREC. In Petr Sojka, Thierry Bouche. DML 2011: Towards a Digital Mathematics Library. Brno: Masaryk University, 2011. p. 77–84. ISBN 978-80-210-5542-1.


FORMÁNEK, David, Martin LÍŠKA, Michal RŮŽIČKA and Petr SOJKA. Normalization of Digital Mathematics Library Content. CEUR Workshop Proceedings, Aachen, 2012, vol. 921, October, p. 91–103. ISSN 1613-0073.


LÍŠKA, Martin, Petr SOJKA, Michal RŮŽIČKA and Peter MRAVEC. Web Interface and Collection for Mathematical Retrieval : WebMIaS and MREC. In Petr Sojka, Thierry Bouche. DML 2011: Towards a Digital Mathematics Library. Brno: Masaryk University, 2011. p. 77–84. ISBN 978-80-210-5542-1.


ŘEHŮŘEK, Radim and Petr SOJKA. Software Framework for Topic Modelling with Large Corpora. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010. p. 46–50. ISBN 2-9517408-6-7.