SIMILARITY SEARCH FOR MATHEMATICS

MICHAL RŮŽIČKA, PETR SOJKA AND MARTIN LÍŠKA

FINE-TUNING QUERY EXPANSION AND UNIFICATION STRATEGIES

INTRODUCTION

Masaryk University (MU) has entered the area of MIR during the development of the Czech Digital Mathematics Library (DML-CZ). It quickly became clear that Digital Mathematical Libraries (DMLs) are specific especially in handling of formulae.

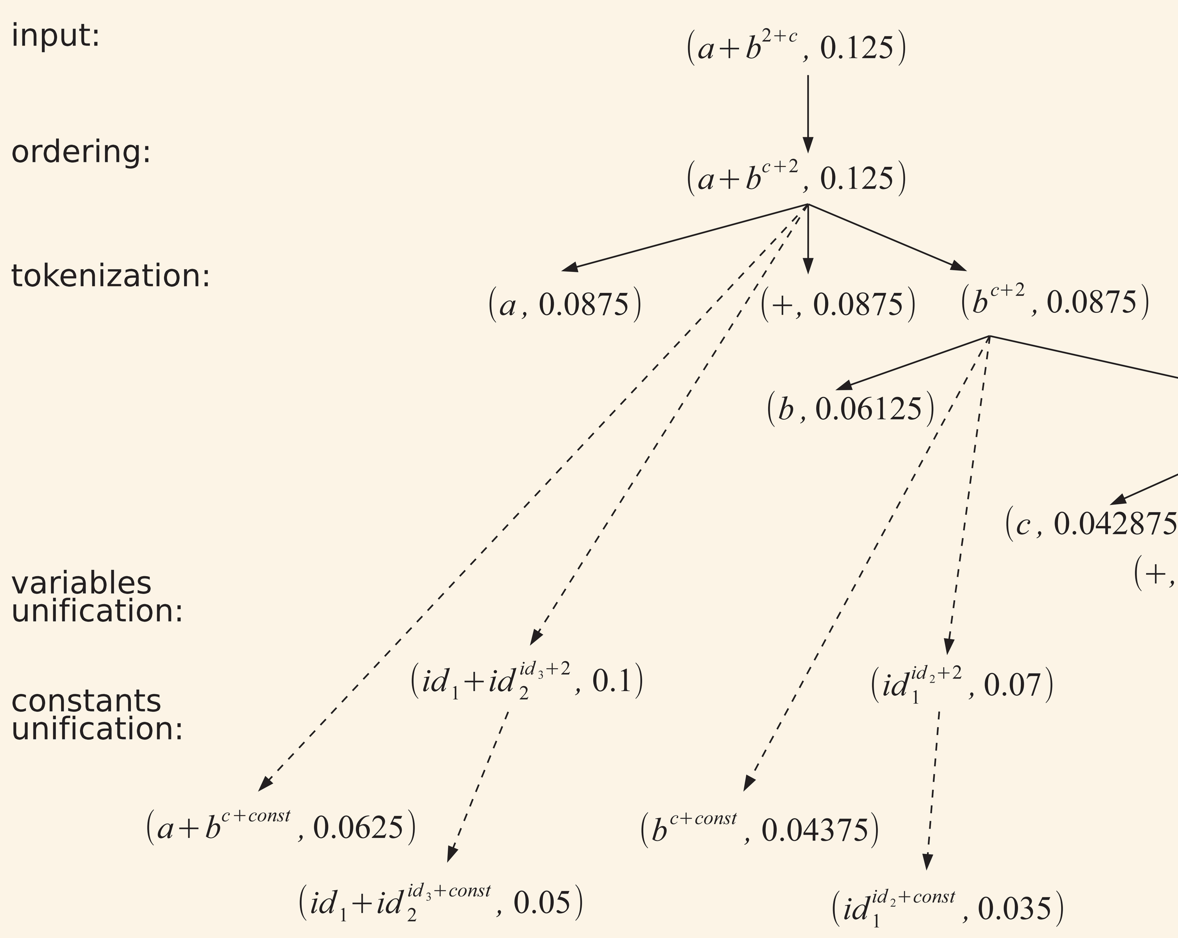
MU has partnered in the development of the European Digital Mathematics Library (EuDML) and supports math formulae search as one of the math specific features. We have also paid attention to the user interface aspect: formulae in the query are rendered at the same time as the user writes it.

EuDML with Math Indexer and Searcher (MIaS) is the first digital library collecting non-born-digital PDFs that supports math search in full texts.

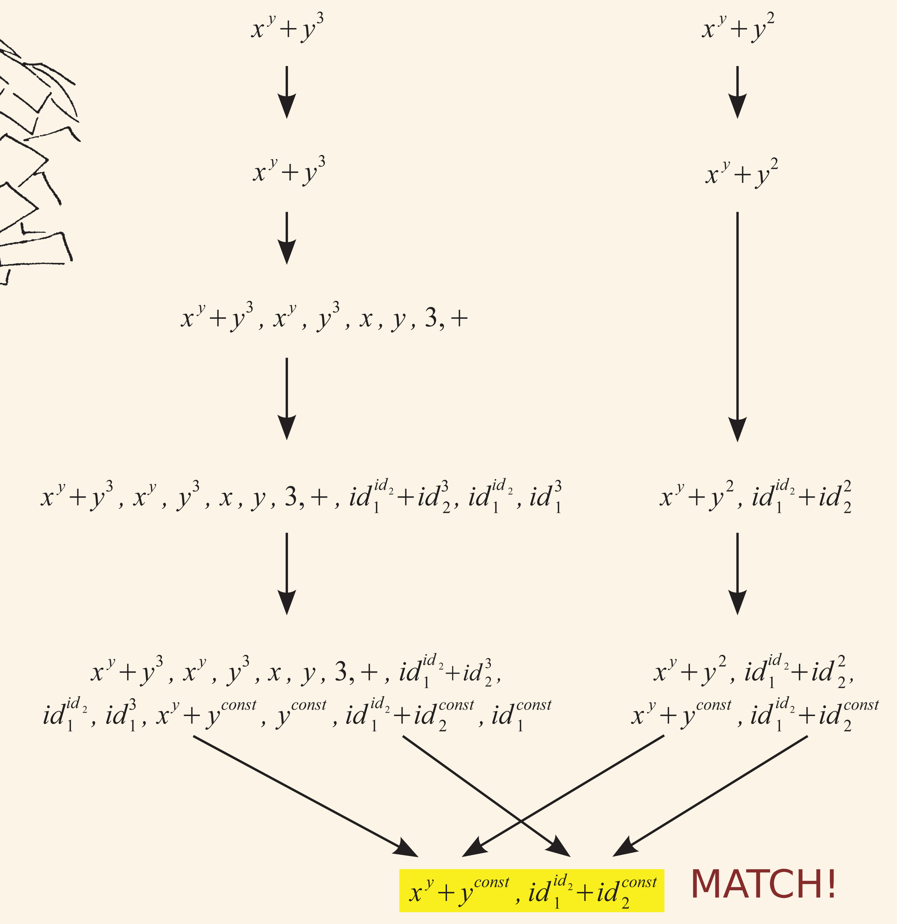
Our MIRMU team has been participating in NTCIR math information retrieval tasks since their introduction at NTCIR-10. This year we have tried three new approaches: structural unification, new querying strategies and new canonicalization procedures.

In our research we found out that structural unification increases recall but has negative impact on precision. NTCIR ground truth allowed us to compare effects of canonicalization and different querying strategies.

MATH PROCESSING



WORKFLOW EXAMPLE



NEW STRUCTURAL UNIFICATION

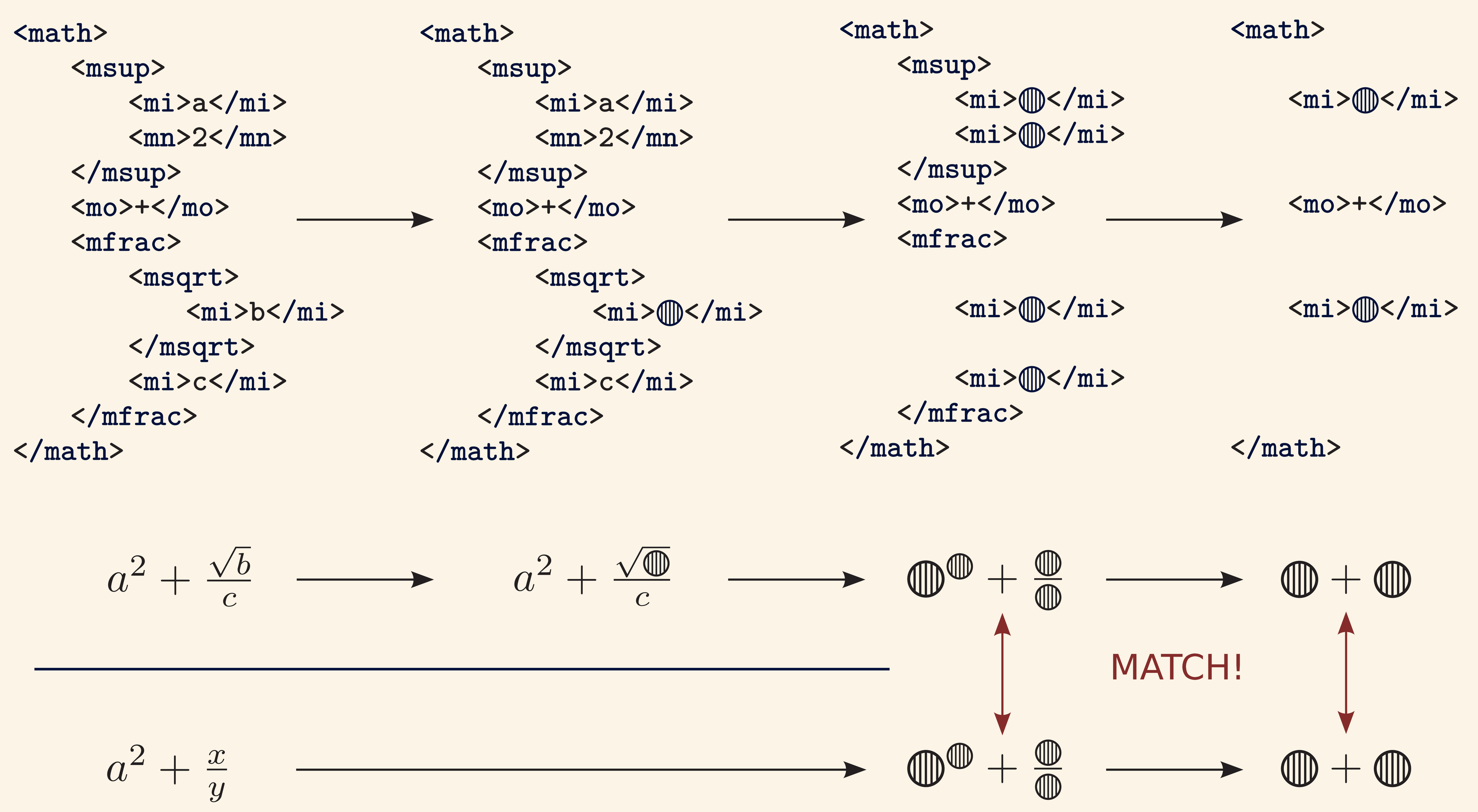
An important feature that we missed in NTCIR-11 was the ability to substitute structures. To structurally unify we implemented the open-source tool MathML Unificator (see https://mir.fi.muni.cz/mathml-normalization/) usable as a standalone command line utility or a Java library embeddable in other systems.

This tool is used for generating structurally unified versions of the input formulae. The unification is performed according to MathML tree layers.

The MathML Unificator tool is now integrated with our (Web)MIaS system and structural unification is done during indexing of formulae of the input documents as well as at query processing to structurally unify formulae from the user queries.

Proper weighting of structurally unified derivatives of the input formulae in relation to the original non-modified formulae and tokenized and unified subformulae is yet to be solved.

Unification is done from lists to the root of the MathML tree of the formula so that the substitution takes place for all the nodes in the given layer in one step and follows layer by layer up to the root. Final unification—to a single ◍—is omitted.



NEW QUERYING STRATEGIES

Based on NTCIR-11 ground truth of annotated data we developed an evaluation framework that allows us to rigorously compare several new querying strategies.

The query relaxation strategy used at NTCIR-11 we call **Leave Rightmost Out (LRO)**:

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

Based on this concept we evaluated also other querying strategies:

Original Query Only (OQO) The basic reference querying strategy is to use the original query without any modifications or derived subqueries.

Math Terms Only (MTO) The query only consists of all formulae from the original query.

Text Terms Only (TTO) In Text Terms Only strategy the query only consists of text keywords from the original query.

All Possible Subqueries (APS) The opposite extreme to OQO is to use all potential subqueries derivable from the original query.

Leave One Out (LOO) This querying strategy is similar to the APS strategy with the following differences:

- We work with a restricted set of the subqueries—only the original query and derived subqueries with exactly one excluded component (one formula or one text keyword).
- Weight of interleaving 'strips' of results from subqueries is 2 if results are taken from the original query results list, and 1 otherwise.

Leave One or Two Out (LOoTO) The Leave One or Two Out querying strategy is a further extension of the previous Leave One Out strategy:

- The set of subqueries consists of the original query and derived subqueries with exactly one or two components excluded.
- The strip-weight is 3 if results are taken from the original query results list, 2 if results are taken from a derived query with exactly one excluded component, and 1 otherwise.

PHRASE EXPANSION

These modifications are only applicable on multi-word text keywords of the original query.

Original query

Formula 1: $\aleph_0$
Keyword 1: categorical simple theory

Phrase expansion For multi-word keywords individual words are used instead of the original multi-word keywords.

Formula 1: $\aleph_0$
Keyword 1: categorical
Keyword 2: simple
Keyword 3: theory

Full phrase expansion Individual words from the multiword keywords are added one by one at the end of the keywords list (removing duplicates, if any).

Formula 1: $\aleph_0$
Keyword 1: categorical simple theory
Keyword 2: categorical
Keyword 3: simple
Keyword 4: theory

This modified version provides the querying strategy more flexibility for query relaxation and boolean operations on the query components.

NEW FEATURES

Canonicalization The canonicalization process aims to normalize potential serializations (different notations in MathML encoding) of the same math formulae. The normalization is optimized for similarity search not to preserve full semantic information of the original formulae but possibly removes semantically negligible differences in behalf of similarity matches.

Canonicalization operators removal List of math operators to be removed from canonicalized formulae:

- U+2062 INVISIBLE TIMES
- U+22C5 DOT OPERATOR
- U+002A ASTERISK
- U+2063 INVISIBLE SEPARATOR
- U+2064 INVISIBLE PLUS

Unary operators removal Unary operators are removed from the input formulae in the process of formulae normalization by our MathML Canonicalizer.

Operator unification We define an operator equivalence relation. We substitute all of these operators with a canonical operator that represents each equivalence class.

Structural unification Indexing structurally unified derivatives of the original formulae was used for the first time by MIaS system at NTCIR-12.

EVALUATION

We have built several dozen indices for the Main and Wiki Math Task with different features and configurations enabled. Every row in the table corresponds to the particular combination of features.

We queried each index with all 50 topics from NTCIR-11 with 11 different querying strategies (columns in the table). This gave us 660 results describing the performance of each particular combination. We used MAP and Bpref metrics for evaluation against NTCIR-11 ground truth.

For NTCIR-12 submission we have chosen four most promising or curious configurations in terms of Bpref and MAP.

Cell colours indicate groups of comparable combinations with respect to Bpref metric.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	BPREF	PCM-T	CM-APSWMWWS	CM-LOOTO	CM-LOO	CM-N1MRIR	CM-N1MRMSFR	CM-N1MRMSF	CM-N1MR	CM-N1M	CM-OQFPO	CM-OQKPO	CM-OQO
4	ntcir-12-10-pfe	0,0492	0,1029	0,1011	0,0987	0,1047	0,1047	0,1057	0,1027	0,1055	0,0220	0,0619	0,0343
5	ntcir-12-11	0,0557	0,0897	0,0873	0,0898	0,0972	0,0977	0,0973	0,0956	0,0977	0,0216	0,0514	0,0348
6	ntcir-12-11-fe	0,0529	0,0964	0,0991	0,1040	0,1105	0,1104	0,1107	0,1079	0,1110	0,0216	0,0673	0,0341
7	ntcir-12-11-pfe	0,0509	0,1037	0,1015	0,0987	0,1045	0,1038	0,1055	0,1019	0,1051	0,0216	0,0616	0,0348
8	ntcir-12-12	0,0552	0,0893	0,0866	0,0891	0,0970	0,0971	0,0969	0,0951	0,0969	0,0226	0,0511	0,0367
9	ntcir-12-12-fe	0,0527	0,0949	0,0969	0,1023	0,1094	0,1090	0,1097	0,1073	0,1094	0,0226	0,0670	0,0351
10	ntcir-12-12-pfe	0,0502	0,1009	0,0989	0,0976	0,1040	0,1033	0,1045	0,1015	0,1040	0,0226	0,0614	0,0357
11	ntcir-12-13	0,0561	0,0882	0,0860	0,0889	0,0960	0,0967	0,0962	0,0945	0,0970	0,0224	0,0513	0,0373
12	ntcir-12-13-fe	0,0529	0,0963	0,0981	0,1032	0,1089	0,1096	0,1094	0,1067	0,1099	0,0224	0,0671	0,0363
13	ntcir-12-13-pfe	0,0517	0,1014	0,1012	0,0986	0,1035	0,1034	0,1043	0,1013	0,1045	0,0224	0,0616	0,0368
14	ntcir-12-14	0,0548	0,0871	0,0858	0,0872	0,0944	0,0945	0,0947	0,0927	0,0947	0,0221	0,0511	0,0356
15	ntcir-12-14-fe	0,0517	0,0935	0,0954	0,1008	0,1066	0,1071	0,1078	0,1046	0,1075	0,0221	0,0668	0,0340
16	ntcir-12-14-pfe	0,0503	0,0995	0,0988	0,0965	0,1021	0,1014	0,1029	0,0996	0,1024	0,0221	0,0613	0,0350
17	ntcir-12-15	0,0965	0,0892	0,0870	0,0895	0,0984	0,0986	0,0984	0,0966	0,0987	0,0220	0,0515	0,0340
18	ntcir-12-15-fe	0,1067	0,0942	0,0980	0,1033	0,1096	0,1102	0,1102	0,1081	0,1106	0,0220	0,0671	0,0332
19	ntcir-12-15-pfe	0,1021	0,1013	0,1001	0,0979	0,1044	0,1040	0,1052	0,1020	0,1048	0,0220	0,0614	0,0338
20	ntcir-12-16	0,0955	0,0884	0,0870	0,0895	0,0974	0,0976	0,0980	0,0961	0,0974	0,0225	0,0512	0,0347
21	ntcir-12-16-fe	0,1058	0,0961	0,0988	0,1038	0,1097	0,1100	0,1104	0,1079	0,1104	0,0225	0,0668	0,0340
22	ntcir-12-16-pfe	0,1017	0,1032	0,1013	0,0989	0,1042	0,1040	0,1051	0,1019	0,1046	0,0225	0,0614	0,0345
23	ntcir-12-17	0,0665	0,0662	0,0682	0,0677	0,0653	0,0639	0,0671	0,0642	0,0659	0,0085	0,0512	0,0411
24	ntcir-12-17-fe	0,0579	0,0602	0,0646	0,0661	0,0616	0,0594	0,0597	0,0578	0,0590	0,0085	0,0670	0,0388
25	ntcir-12-17-pfe	0,0570	0,0685	0,0708	0,0706	0,0653	0,0654	0,0662	0,0640	0,0674	0,0085	0,0615	0,0440
26	ntcir-12-18	0,0467	0,0438	0,0443	0,0446	0,0433	0,0419	0,0434	0,0420	0,0427	0,0017	0,0512	0,0236
27	ntcir-12-18-fe	0,0495	0,0484	0,0509	0,0529	0,0509	0,0484	0,0479	0,0463	0,0481	0,0017	0,0668	0,0285
28	ntcir-12-18-pfe	0,0496	0,0496	0,0540	0,0540	0,0477	0,0459	0,0470	0,0458	0,0466	0,0017	0,0613	0,0273
29	ntcir-12-19	0,0000	0,0697	0,0699	0,0681	0,0688	0,0681	0,0702	0,0679	0,0696	0,0089	0,0514	0,0361
30	ntcir-12-19-fe	0,0000	0,0763	0,0781	0,0794	0,0800	0,0772	0,0771	0,0748	0,0772	0,0089	0,0673	0,0432
31	ntcir-12-19-pfe	0,0000	0,0770	0,0809	0,0794	0,0757	0,0734	0,0745	0,0729	0,0745	0,0089	0,0617	0,0417

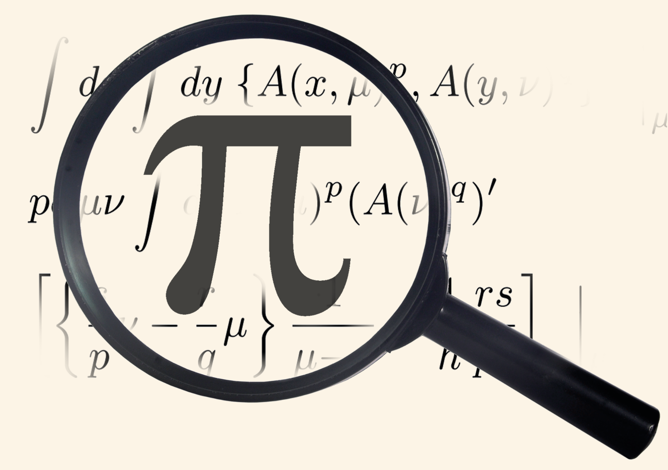
CONCLUSIONS

The main advances of our approach since NTCIR-11 Math Task were development of our evaluation platform based on NTCIR-11 ground truth and introduction of math structural unification component as a part of the MIaS processing workflow.

Use of structurally unified derivatives increases recall but has negative impact on precision. Fine tuning the weights of structural unification nodes could possibly balance performance of our system towards recall at the expense of precision and vice versa. Setting and tuning the indexing and preprocessing parameters is necessary for given application.

We aim to reuse NTCIR-12 MathIR data as the ground truth in our evaluation platform to further improve performance of our system.

Our future MathIR research aims at incorporating machine learning techniques to formulae disambiguation and ranking, and deploying Computer Algebra Systems for better canonicalization of the input formulae.



Contact Information
mruzicka@mail.muni.cz, sojka@fi.muni.cz, martin.liski@mail.muni.cz, Faculty of Informatics, Botanická 68a, 602 00 Brno, Czech Republic

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

Citation
RŮŽIČKA, Michal; SOJKA, Petr; LÍŠKA, Martin. Math Indexer and Searcher under the Hood: Fine-Tuning Query Expansion and Unification Strategies. In Noriko Kando, Kazuaki Kishida, Makoto P. Kato, and Shuhei Yamamoto. Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan, 2016. p. 331-337, 7 pp. ISBN 978-4-86049-071-3. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/ntcir/MathIR/05-NTCIR12-MathIR-RuzickaM.pdf

https://is.muni.cz/publication/1346605?lang=en

Publications

LÍŠKA, Martin; SOJKA, Petr; RŮŽIČKA, Michal. Combining Text and Formula Queries in Math Information Retrieval: Evaluation of Query Results Merging Strategies. In: Proceedings of the First International Workshop on Novel Web Search Interfaces and Systems. NWSearch'15. Melbourne, Australia: ACM, 2015, p. 7–9. ISBN 978-1-4503-3789-2. doi:10.1145/2810355.2810359.

RŮŽIČKA, Michal; SOJKA, Petr; LÍŠKA, Martin. Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy. In Hidei Joho, Kazuaki Kishida. Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2-1-2 Hitotsubashi, 2014. 8 pp. ISBN 978-4-86049-065-2.

SOJKA, Petr; LÍŠKA, Martin. 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, 4 pp. ISBN 978-1-4503-0863-2. doi:10.1145/2034691.2034703.

SOJKA, Petr; LÍŠKA, Martin. Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In James H. Davenport, William M. Farmer, Josef Urban, Florian Rabe. Intelligent Computer Mathematics Lecture Notes in Computer Science, 2011, Volume 6824/2011. Berlin / Heidelberg: Springer, 2011. p. 228–243. ISBN 978-3-642-22672-4. doi:10.1007/978-3-642-22673-1_16.

LÍŠKA, Martin; SOJKA, Petr; RŮŽIČKA, Michal; MRAVEC, Peter. 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.

Illustrations by Jiří Franek