European Science Foundation, Research Networking Programmes
ESF Short Visit Grant 6967 Scientific Report

Petr Sojka

Contents

1 Purpose of the visit (proposal title): Evaluation of Math Indexer and Searcher (MIaS) for Math Information Retrieval (MIR)

Aim of the visit was to present and discuss evaluation of the MIaS, the Math Indexer and Searcher system, that has been developed at Masaryk University (MU) since 2008, and get new directions at further developments. Our MIR team at MU (MIRMU) registered for the second Math task (Math-2) at the NTCIR-11 (Evaluation of Information Access Technologies) conference held at the National Institute of Informatics, Tokyo, Japan (`http://research.nii.ac.jp/ntcir/ntcir-11/conference.html`). It was a unique opportunity to compare our MIR research, approaches, engine and evaluation results with leading experts in the field, which are working specifically in the Math Information Retrieval domain, and with participants of the 6th International Workshop on Evaluating Information Access (EVIA 2014), a collocated satellite workshop of the NTCIR-11 conference.

Hosting teams of prof. Noriko Kando and prof. Akiko Aizawa organized Math-2 task related program of the conference:

- Wikipedia Subtask pre-conference meeting
  Date: December 8, 2014
  Place: NII Seminar Rm. 2005, floor 20
  Web: `http://ntcir11-wmc.nii.ac.jp/index.php/NTCIR-11-Math-Wikipedia-Task#Schedule`
- Task Overview 4: Math-2
  Date: December 10, 2014
  Web: `http://research.nii.ac.jp/ntcir/ntcir-11/program.html`
- Math-2 Oral Session
  December 11, 2014
  Place: Room 4
  Web: `http://research.nii.ac.jp/ntcir/ntcir-11/program.html`
- Math-2 Poster Session
  December 11, 2014
  Place: Room 2 & 3
  Web: `http://research.nii.ac.jp/ntcir/ntcir-11/program.html`
- Math-2 Round Table Session
  December 11, 2014
  Place: NII Seminar Rm. 2006, floor 20
  Web: `http://research.nii.ac.jp/ntcir/ntcir-11/program.html`

I have attended and took advantage of participation and discussions at the all Math-2 related events held at NII from December 8th to December 11th and selected events of other tasks.

We also attended EVIA workshop, as seen on Figure 1 on the following page.

2 Description of the work carried out during the visit

I have attended all Math-2 related events and actively participated in them.

At pre-conference Wikipedia task meeting we discussed our results in the new Math subtask of NTCIR-11 with the main organizer Moritz Schubotz (TU Berlin, Germany) and other participants.

During the main conference program we focused on the new strategies used for Math-2 task at NTCIR-11 this year. At the round table session at the end of the conference we discussed how to push forward the frontiers of evaluation of mathematics retrieval with prof. Noriko Kando, NTCIR-11 programme co-chair, prof. Akiko Aizawa, NTCIR-11 Math-2 task organizer and experts in Math NLP techniques at NII, namely with Michael Kohlhase (Jacobs University Bremen), Iadh Ounis (University of Glasgow), Moritz Schubotz (TU Berlin, Germany), Richard Zanibbi (Rochester Institute of Technology, USA) and others.

With my doctoral student Michal Růžička, we

Figure 1: Petr Sojka attending EVIA conference

- presented results of our team in Math-2 Wikipedia subtask during the pre-conference Wikipedia subtask meeting (NII Seminar Rm. 2005, floor 20, December 8 14:00–18:00), together with the broader vision of further research and development in the MIR domain. Slides are available: `https://is.muni.cz/publication/1212774/en`
- presented paper *RŮŽIČKA, Michal, Petr SOJKA and Martin LÍŠKA. Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy. In Noriko Kando, Hideo Joho, Kazuaki Kishida. Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan, 2014. p. 127–134, 8 pp. ISBN 978-4-86049-065-2.* as part of Math-2 Oral Session (December 10 14:40–15:10): `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/Math-2/07-NTCIR11-MATH-RuzickaM.pdf`
- presented poster *Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy* at Math-2 Poster Session (December 10 12:35–14:05): `https://www.fi.muni.cz/usr/sojka/posters/ruzicka-sojka-liska-ntcir2014.pdf`
- took part at Math-2 Round Table Session (NII Seminar Rm. 2006, floor

20, December 11 16:05–18:00) and helped to asses this year's Math-2 task and form next year's setup.

3 Description of the main results obtained

We have evaluated MIaS as a promising system with high recall. Our MIaS actually won the Math-2 task of NTCIR-11 as the maths-aware search system with the best results in most evaluated metrics (e.g. ranked #1 with bpref metrics on all subtasks). Runs with Content Math slightly outperformed the ones based on Presentation MathML. Our system implemented query expansion strategy that melted textual and math hits in the best way.

We have realized that further developments and directions are still possible:

1. formulae unification during indexing and querying to increase recall;
2. MathML canonicalization and semantic annotation additions (e.g. linking named entities to math formulae) to increase precision;
3. implementation of Presentation to Content MathML conversion with disambiguation of formulae semantic markup: using of NLP, machine translation, machine learning techniques seems necessary;
4. develop ground truth for testing our engine to implement evaluation driven development;
5. evaluate possibility of metric similarity search indexing to have even better recall.

4 Future collaboration with host institution (if applicable)

We plan to continue further research and joint cooperation and participation at NTCIR-12 (Math-3 task, if any), and discussed eventual MOU and granting possibilities in MIR evaluation within Masaryk University and other European, American or Japanese institutions. We evaluate NII International exchange activities `http://www.nii.ac.jp/en/about/international/` especially NII International Internship Program `http://www.nii.ac.jp/en/about/international/mouresearch/` for Ph.D. students' exchange.

5 Projected publications/articles resulting or to result from your grant

RŮŽIČKA, Michal, Petr SOJKA and Martin LÍŠKA. Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy. In Noriko Kando, Hideo Joho, Kazuaki Kishida. Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo:

National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo
101-8430 Japan, 2014. p. 127–134, 8 pp. ISBN 978-4-86049-065-2.

> This paper describes and summarizes experience of Masaryk University Math Information Retrieval team (MIRMU) with the mathematical search developed and performed for the NTCIR-11 Math-2 Task. Our approach is the similarity search based on canonicalized MathML and second generation of scalable full text search engine Math Indexer and Searcher (MIaS) with attested state-of-the-art information retrieval techniques like query expansion. The capability of MIaS system in terms of math query notation, normalization and combining math with textual query tokens was deployed by submitting multiple runs with four query notations provided, and with results merged from multiple queries. The analysis of the evaluation results shows that the system performs best using TeX queries that are translated and canonicalized to Content MathML, where MIaS ranked as #1 for all metrics returning very relevant results.

For further details, see `https://is.muni.cz/publication/1201956/en`

We expect to participate at NTCIR-12 and publish our MIR results there, and at CICM 2015 conference in Washington DC, USA, or at SIGIR 2015.

6 Other comments (if any)

We thank for the perfect organization and hospitality of NII, especially prof. Akiko Aizawa, and are grateful for the ELIAS support.