Exploring the One-brain Barrier: a Manual Contribution to the NTCIR-12 MathIR Task
Moritz Schubotz, Norman Meuschke, Marcus Leich, and Bela Gipp

Math Information Retrieval
- Formulae: integral part of language in STEM
- Information Retrieval
- Today: Textual content
  - Recently: Image and Video
- Exponential growth in publications
- Applications
  - Applicable Theorem Search
  - Plagiarism Detection
  - Related work search
  - Math IR needs
- Datasets with structured mathematical formulae
- Topics
- Evaluation methods

Motivation
- Formulae: integral part of language in STEM
- Information Retrieval
- Today: Textual content
  - Recently: Image and Video
- Exponential growth in #publications
- Applications
  - Applicable Theorem Search
  - Plagiarism Detection
  - Related work search
  - Math IR needs
- Datasets with structured mathematical formulae
- Topics
- Evaluation methods

Example result
Query
Difference between Log \( x_1 \) and log \( x_1 \)

Conclusion
- Gold standard dataset as big step forward to develop a math aware search engine for Wikipedia
- System strengths:
  - Definition lookup queries
  - Applications lookup
- System weaknesses
  - Low precision
  - No standard interface to specify query type

Future work
- Improve description of information need
  - taking into account our focussed mir task categories
  1. Definition look-up
  2. Explanation look-up
  3. Proof look-up
  4. Application look-up
  5. Computation assistance
  6. General formula search
- Improve query syntax for similarity search
- Develop a math search engine for wikipedia with the help of the gold standard dataset and new Mathematical Language Processing Technology

Contact
Moritz Schubotz (Uni Konstanz)
schubotz@tu-berlin.de
+49 7531 88 4438
(Mobile: +49 1578 047 1397)
www.formulasearchengine.com

10.06.2016
www.formulasearchengine.com