Effective Indexing, Querying and Searching

in Math Information Retrieval

Dávid Lupták
dluptak@mail.muni.cz

Faculty of Informatics, Masaryk University

November 25, 2021

Introduction

Math Information Retrieval

- How to involve math-awareness of science, technology, engineering and mathematics (STEM) fields?
- Conventional search engines cannot handle mathematics
  - e.g. `$E=mc^2$` in LaTeX notation
- Task for digital mathematics libraries (DML)
  - support for math representations (LaTeX, MathML, built-in input, ..) in search

Topics diversity

State of the art

- Topics (and their categories) from ARQMath lab
- TopicEq: A Joint Topic and Mathematical Equation Model for Scientific Texts
- Word / word + math / math embeddings
- Topic classification – then select the best system to solve / search on this topic
- Ensemble systems – commissions, voting strategies
- Computer algebra systems, math solvers

Topics diversity

Aims of the Thesis

- Topics categorized as computation, concept, and proof.
- Topics complexity ranges from easy, medium, to hard.
- Math dependency on the surrounding texts within mixed queries.

- What is the effectiveness of different math-aware system setups for different categories of topics?

Topics diversity

Perspectives

Types of math questions / topics[1]:

- easy – related to computation, dependent on formula
- hard – related to mathematical concept, dependent on text
- normal – related to proof, dependent on both text and formula

- Different indices for different purposes
- Ensemble systems

[1] ARQMath "classification"

Math representations in documents

State of the art

Consider various/different math representations

- tokens as bag of math symbols
- tree-based representation

Math representations in documents

$$a^2 + b^2 = c^2 \quad \text{or} \quad x^2 + y^2 = z^2$$

- Look at the structure!
- We don't care about the variable names
- Based on the structure of the formula, we know what represents

Math representations in documents

$$E = mc^2$$

- Look at the context, domain knowledge!
- We all know what this well known formula means
- We can easily substitute variable names with their meaning, because we have *general* knowledge

Math representations in documents

Aims of the Thesis

$$a^2 + b^2 = c^2 \quad \text{or} \quad x^2 + y^2 = z^2$$

- Formula appearance – visual presentation.
- Formula syntax – mathematical (operational) syntax, semantics.
- Discrete representation of math – math formula as a tree.
- Continuous representation of math – math formula in latent space.

- How to grab and disambiguate the meaning of symbols in the formulas?

Math representations in documents

Perspectives

- Canonicalization & unification
- System setups from the previous evaluation workshops (NTCIR, ARQMath)
- ARQMath-3 new task (Open-Domain Question Answering)

Evaluation background

Datasets and evaluation workshops for math information retrieval

- MREC dataset: ca 439 thousand documents
- Math Stack Exchange: ca 1 million questions, ca 28 million LaTeX formulas
- ArXiv.org: ca 1.8 million documents

- NTCIR 10, 11, 12 (years 2013–2016)
- ARQMath (2020, 2021)

Thank You for Your Attention!

Maths Information Retrieval research group at Masaryk University

`https://mir.fi.muni.cz/`
`https://github.com/MIR-MU`

MUNI
FACULTY
OF INFORMATICS