Math Information Retrieval using a Conventional Search Engine

Frank Wm. Tompa

David R. Cheriton School of Computer Science
University of Waterloo
fwtompa@uwaterloo.ca

1 Introduction

Documents in the STEM disciplines rely heavily on the use of math formulas to express knowledge. Searching a corpus of such documents, therefore, requires that a search engine be effective in matching formulas and math terminology, as well as natural language text.

In our presentation, we describe a simple representation for features extracted from math formulas, our implementation of a prototypical search engine, and performance results against benchmarks created to evaluate math-aware search engines for community question-answering.

This research is being conducted in collaboration with Andrew Kane (PhD 2014, Waterloo).

2 Further Details

Effective math information retrieval has been under investigation by several students who have worked under my direction [3, 1, 5] resulting in a best paper award [2] and a best of labs designation [6]. Notably, this last paper describes how natural language mathematical questions can be automatically transformed into formal queries consisting of keywords and formulas and how the resulting formal queries can be effectively executed against a corpus. A key component of our approach has been to represent each formula as a bag of math features and to treat those features as simple search terms. This approach can be adopted by any conventional, text-based search engine, including one developed recently to explore various aspects of search technology.[1]

We describe the three major processing steps used in our system:

1. query construction (how to convert natural language questions into formal queries by selecting and augmenting the text and formulas),
2. mapping formal queries to search terms (especially, how to choose suitable features to represent math formulas), and
3. indexing and querying with the search engine (how to run queries efficiently and rank results effectively).

We also summarize some of our experimental results from the ARQMath Labs [4], a benchmark based on a collection of questions and answers from Math Stack Exchange (MSE) between 2010 and 2018 consisting of approximately 1.1 million question-posts and 1.4 million answer-posts. The main task presents experimenters with 100 mathematical questions (selected by the organizers from MSE question-posts in a subsequent year) and asks for ranked lists of potential answers among existing answer-posts in the collection.

Finally, we outline what research we intend to undertake to improve each of the three processing steps.

[1] https:// github.com/andrewrkane/mtextsearch

References

[1] Dallas J. Fraser. Math information retrieval using a text search engine. Master's thesis, University of Waterloo, Cheriton School of Computer Science, 2018.

[2] Dallas J. Fraser, Andrew Kane, and Frank Wm. Tompa. Choosing math features for BM25 ranking with Tangent-L. In *DocEng 2018*, pages 17:1–17:10, 2018.

[3] Shahab Kamali. *Querying Large Collections of Semistructured Data.* PhD thesis, University of Waterloo, Cheriton School of Computer Science, 2013.

[4] Behrooz Mansouri, Vit Novotný, Anurag Agarwal, Douglas W. Oard, and Richard Zanibbi. Overview of ARQMath-3 (2022): Third CLEF lab on Answer Retrieval for Questions on Math. In *CLEF 2022*, volume 13390 of *LNCS*. Springer, 2022.

[5] Yin Ki Ng. Dowsing for math answers: Exploring MathCQA with a math-aware search engine. Master's thesis, University of Waterloo, Cheriton School of Computer Science, 2021.

[6] Yin Ki Ng, Dallas J. Fraser, Besat Kassaie, and Frank Wm. Tompa. Dowsing for math answers. In *CLEF 2021*, volume 12880 of *LNCS*, pages 201–212, 2021.