Appearance-Based Retrieval of Mathematical Notation in Documents and Lecture Videos

Kenny Davila
Department of Computer Science
Rochester Institute of Technology
Rochester, New York, USA
kxd7282@rit.edu

ABSTRACT

Large data collections containing millions of math formulae in different formats are available on-line. Retrieving math expressions from these collections is challenging. Based on the notion that visually similar formulas are related, we propose a framework for appearance-based formula retrieval in two different modalities: symbolic for text documents and image-based for videos.

We believe that we can achieve high quality formula retrieval results using the visual appearance of math notation without complex formula semantic analysis.

We represent mathematical notation using different graph types to take advantage of the information available on each domain. For symbolic formula retrieval, math expressions in text formats like LaTeX are parsed to generate Symbol Layout Trees [4]. For image-based formula retrieval, image processing techniques are used to create a graph-based image content representation.

We store these graphs using an inverted index of pairs of primitives defined by the triplet (p, q, r), where p and q are the labels of two primitives connected in the graph by the path r [4].

Retrieval is a two-stage process: candidate selection and reranking. The first stage uses pairs of primitives from the query graph to find matches in the inverted index. Each match is given an initial score using the Dice coefficient of matched pairs of primitives [4].

The best top-K candidates from the first stage are selected for re-ranking using a detailed similarity metric. Two steps are performed for each candidate: matching and scoring. The **matching** step is done by searching for the largest common substructure between query and candidate graphs. Matching is related to the problem of finding the maximum common subgraph isomorphism (MCS) between two graphs. In addition, we consider label unification for symbolic formula retrieval, and our wildcard query nodes can match entire subgraphs. In the **scoring** step, multiple similarity criteria define a score vector used to sort candidates, either by lexicographic order or by a function of these scores.

In the next stage of our project, different datasets and benchmarks will be required to evaluate each modality. For symbolic formula retrieval, we will use the most recent versions of the NTCIR MathIR Tasks benchmarks [1]. To the best of our knowledge, there are no benchmarks for large scale image-based formula retrieval. However, the same collections used for symbolic formula retrieval could be adapted by rendering math expressions to images. In addition, we will use datasets of math lecture videos for image-based formula retrieval.

Traditional graded-scales of relevance used for evaluation of retrieval systems have been shown to have inconsistency issues [2]. We plan to use pairwise candidate comparisons during our evaluation phase. Some aggregation methods exist that generate relevance scores and ideal rankings using these pairwise candidate comparisons [3].

The proposed framework can be adapted to work for other domains like chemistry or technical diagrams where visually similar elements are usually related.

Keywords

Mathematical Information Retrieval, Content-Based Image Retrieval, Graph-Based Retrieval

Acknowledgments

This material is based upon work supported by the National Science Foundation (USA) under Grant No. HCC-1218801

1. REFERENCES

[1] A. Aizawa, M. Kohlhase, I. Ounis, and M. Schubotz. NTCIR-11 Math-2 task overview. In *Proceedings of NTCIR-11 Math-2 task Workshop Meeting [1]*, 2014.

[2] K. Radinsky and N. Ailon. Ranking from pairs and triplets: Information quality, evaluation methods and query complexity. In *WSDM*, pages 105–114, New York, NY, USA, 2011. ACM.

[3] M. N. Volkovs and R. S. Zemel. New learning methods for supervised and unsupervised preference aggregation. *JMLR*, 15(1):1135–1176, 2014.

[4] R. Zanibbi, K. Davila, A. Kane, and F. Tompa. Multi-stage math formula search: Using appearance-based similarity metrics at scale. *SIGIR*, 2016.

SIGIR '16 July 17-21, 2016, Pisa, Italy

ACM ISBN 978-1-4503-4069-4/16/07.

DOI: http://dx.doi.org/10.1145/2911451.2911477