Tangent 1.1: Math and Text Search Engine

Nidhin Pattaniyil (ntp5633@rit.edu) ***Advisor:*** *Richard Zanibbi*
Rochester Institute of Technology

Introduction

Tangent v1.1 is a math/text search engine that allows the user to enter a query containing latex/draw equations and text.

They are two separate indices: Solr for textual terms and a custom index for formulas.

A formula is parsed into a Symbol Layout tree and pairs of symbols found in the tree are stored in an inverted index.

The following corpuses are indexed: Wikipedia and a subset of the arXiv.org

Contributions

Scalable:
Optimized storage of Math index to scale to larger datasets (100 gb+)

Indexing Matrices:
Supported indexing of matrices

Integration of Text:
Added support of indexing text in documents using Solr

Hit Summarization:
Implemented a hit summarization algorithm[3] to summarize hits

Wildcard:
Allow for single symbol wildcard queries

Acknowledgments

The material is based upon work supported by National Science Foundation(USA) under Grant No. IIS-10161815

Ranking Function

Given a query q and a document d, the scoring function is

$$\text{score(q,d)} = 0.6 * ms(q,d) + 0.3 * ts(q,d) + 0.1 * ps(q,d)$$

where
ts is the similarity of the text terms in the query and document; the score is provided by Solr and normalized to 1; similarity is calculated through tf-idf

ps: is the similarity of the positions of the terms in the query compared to a document. Pairs of terms that appear together are preferred

ms is the similarity of the formula in the query and in the document. Similarity is defined as the number of symbols that matched in a query expression and an expression in a document

Math Tokenization

For the expression **z=(5)+1,** the tokens generated are

matirx mstart|mend|1|1, matrix_1_1|1|2|0 ,
matrix_1_1|+|1|0
matrix|5|0|0, 5|None|0|0, 1|None|0|0
z|+|3|0 , z|=|1|0, z|1|4|0 , z|matrix_1_1|2|0
=|matrix_1_1|1|0 ,=|1|3|0, =|+|2|0,
+|1|1|0

Expressions with wild cards

Supports wildcard in math expression that unify to single math symbols

$$a^{?i} + b^{?i} = c^{?i}$$

?i can match numbers such as 99, or identifiers z

Evaluation

Our system was compared against another system ,MIAS[2] with the MREC dataset.

MREC dataset mostly research papers from subject like chemistry, physics

Queries containing wildcard queries, matrices, text and expression were generated based on scanning the dataset.

$$\textbf{Wildcard Query}: a^{?n} + b^{?n} = c^{?n}$$

$$\textbf{Matrix Query}: \begin{pmatrix} \alpha+1 & & \\ & \alpha+1 & \\ & & \alpha+1 \end{pmatrix}$$

$$\textbf{Full Query}: \text{Fermat's little theorem } a^P \equiv a(mod P)$$

Conclusion

Comparison between MIAS shows our systems returns the most relevant result on the top compared to MIAS for all types of queries

Similarity of the expression in a query and a document best captures if a document is relevant

Evaluating partial relevance is difficult for an unfamiliar domain

Future Work

Experiment with different scoring parameters

Compare system with different implementations

Improve the speed of indexing and retrieving results

Support wildcard queries in the user interface

References

[1] D. Stalnaker (2013) Math Expression Retrieval Using Symbol Pairs in Layout Trees
[2] Martin Líška,Petr Sojka. Evaluation of Mathematics Retrieval
[3] Varadarajan,Ramakrishna. A system for query-specific document summarization