MathFind: A Math-Aware Search Engine

Rajesh Munavalli
Design Science, Inc.
140 Pine Ave, 4th Floor
Long Beach, CA 90802
+1-651 223-2884
rajeshm@dessci.com

Robert Miner
Design Science, Inc.
140 Pine Ave, 4th Floor
Long Beach, CA 90802
+1-651 223-2883
robertm@dessci.com

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Abstracting methods, Indexing Methods, Linguistic Processing;*

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query formulation, Retrieval models, Search process.*

General Terms

Algorithms, Design.

Keywords

Math Search, MathML, Equation Editor.

1. SYSTEM OVERVIEW

Researchers working in technical disciplines wishing to search for information related to a particular mathematical expression cannot effectively do so with a text-based search engine unless they know appropriate text keywords. To overcome this difficulty, we demonstrate a math-aware search engine, which extends the capability of existing text search engines to search mathematical content**.**

Our search engine is composed of a MathFind processing layer implemented on top of a typical text-based search engine layer. Our prototype piggybacks upon the Apache Lucene Search API, a modified vector space model-based text retrieval system.

The MathFind layer of the search engine analyzes expressions in MathML, an XML standard for representing mathematical notation. The process decomposes the mathematical expression into a sequence of text-encoded math fragments. These math fragments are analogous to words in a text document. Math fragments combined with text content serve as input to the text-search engine. At query time, a graphical equation editor is used to enter a math query, which is internally represented in MathML. The math-processing layer converts the MathML query into a sequence of text-encoded math query terms, which form the basis of a text query performed by the underlying text-search engine. To overcome the ambiguity in the presentation of an expression, MathML input is normalized before processing. [1, 2]

The current implementation has the following features

- Indexes variety of document formats: text + MathML, XHTML + MathML, DocBook + MathML, and via conversion, LaTeX, MS Word, and Mathematica notebooks.
- The search engine retrieves ranked documents based on similarity to both math and text queries.
- The system is capable of interpreting wild card queries in math expressions analogous to text wild queries.
- Math query terms can be highlighted in the retrieved documents (cached) to help users locate matched expressions.

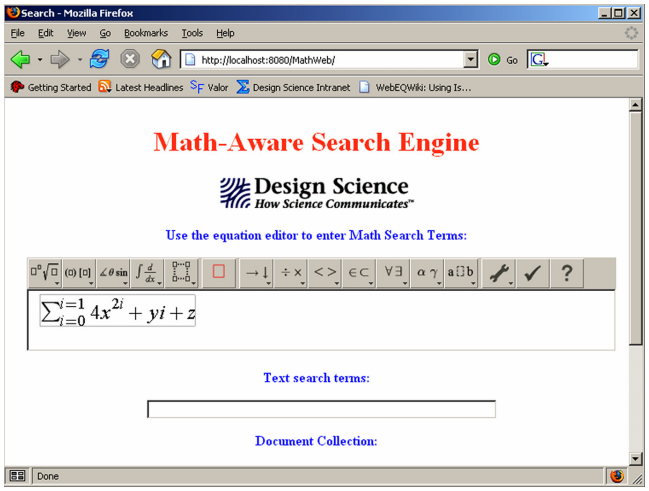

Figure 1: Search Engine Query Interface

2. ACKNOWLEDGEMENTS

This work is supported in part by the National Science Foundation through the National Science Digital Library program under Grant No. 0333645. We would also like to thank Dr. Abdou Youssef of GWU for helpful discussions.

3. REFERENCES

[1] Salton, G., Fox, E., Wu, H. Extended Boolean information retrieval. *Communication of the ACM, 26(11):1022-1036, 1983.*

[2] Ogilvie, P., and Callan, J. Using Language models for flat text queries in XML retrieval. *Proceedings of INEX 2003 Pages12-18.*

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.