System Description: EgoMath2 As a Tool for Mathematical Searching on Wikipedia.org*

Jozef Mišutka[1] and Leo Galamboš[2]

[1] Department of Software Engineering, Charles University in Prague
misutka@ksi.mff.cuni.cz
[2] Cythres group, Czech Technical University in Prague
galamleo@fd.cvut.cz

1 Introduction

EgoMath is a full text search engine focused on digital mathematical content with little semantic information available. Recently, we have decided that another step towards making mathematics in digital form more accessible was to enable mathematical searching in one of the world's largest digital libraries - Wikipedia. The library is an excellent candidate for our mathematical search engine because the mathematical notation is represented by TEX fragments which do not contain semantic information.

The key issue in mathematical searching is the retrieval of mathematically equal formulae. We regard this issue as a similarity search problem where the similarity function is strongly dependent on the mathematical model. EgoMath2 and its predecessor use the same idea but different implementation for the similarity function, indexing and searching. The idea is to use content-based annotation - different textual representations of one formula which are mathematically similar in our definition - for allowing similarity search. The similarity of mathematical formulae in EgoMath2 is based on the mathematical equality in a predefined mathematical model and preferring operations to operators in their formula parse trees. The later characteristic is used in generalising the formula representation e.g. formula $a + 7$ is generalised to $id + 7$. Currently used model consists of several rules e.g. commutative property of addition. One possible definition of similarity can be found in [1]. The similarity search is hidden in multiple queries which can be performed while searching for one formula in the space of equal and similar representations in the index.

A textual representation of a mathematical formula is a TEX-like flattened symbol representation by words e.g. a^2 is represented by three words: a, ^, 2, internally represented in postfix notation to avoid parenthesis issues. There are two important algorithms used during the content-based annotation. The *augmentation* algorithm exploits the biggest advantage of full text search engines - fast searching in a huge set of words. Consequently, for each formula the algorithm produces several representations consisting of ordered words which are

* This work was supported by the grant SVV-2011-263312.

J.H. Davenport et al. (Eds.): Calculemus/MKM 2011, LNAI 6824, pp. 307–309, 2011.

indexed like normal text. Each textual representation of a formula is equal to or less similar than the previous one.

The *ordering* algorithm converts each representation to a canonical one. The ordering algorithm guarantees that two mathematically equal formulae with the same but permuted operands have the same unique canonical representation. Two similar (but not equal) formulae have a similar textual representation.

Semantically rich mathematical formulae cannot be represented by a full text search engine without losing the semantic information in general. The augmentation and ordering try to minimise this disadvantage.

2 New Features and Architecture Changes in EgoMath2

EgoMath2 is based on the newest version of the Java full text search engine Egothor (http://www.egothor.org). Strongly decoupled architecture of the mathematical extension and the full text indexer made the update straightforward. Learning from our experience with the first version, the augmentation process in EgoMath2 was made easily extensible and configurable using XML configuration files. Both the algorithms which are applied during the augmentation and the ambiguous symbol meaning can be configured to take advantage of additional knowledge about the underlying document set. Architectural changes were made allowing for ranking the query results.

The graphical user interface(UI) had been completely rewritten [2], thus the mathematical support had to be implemented from scratch. The connection between indexer and the UI was simplified, a new text element for mathematical input was added, snippets showing matched formula representations were introduced and debugging capabilities were also improved. The new UI has administration features which can be useful in online mathematical systems. EgoMath2 supports roles with different privileges offering different search indexes. The indexer web administration has proven itself useful for quick document set inspection. The performance of the mathematical indexer was improved (EgoMath2 is 3x faster in indexing Wikipedia dataset than the previous version) by caching formulae string representations and by small optimisations and cleanup of code.

3 Adjusting Wikipedia for EgoMath2 and vice versa

Preparing Wikipedia for indexing by EgoMath2 means to download articles from Wikipedia.org, sort out non-mathematical articles, convert mathematical notation into supported format and create HTML pages which are fed to EgoMath2. A dump of English articles from January 2011 (30GB) was downloaded from the official website [3]. One by one, all types of articles were extracted. The mathematical articles were identified by looking for the string "<math>". $28,376$ mathematical articles (425MB dump) have been found with more than $240,000$ mathematical elements from more than 10 million articles. We tried to convert each element to semantically richer MathML using latex2mathml web

service developed by the KWARC group [4] to improve the semantic information. EgoMath2 then uses both the LaTeX and the MathML format of the formula if available. More than 300 new symbols (e.g. Invisible Separator U+2063) have been added into our XML symbol configuration to improve the semantic quality of the extracted mathematical formula. Several modifications had to be made because of incorrect conversions of complex formulae. The error checking had to be relaxed. The parser heuristics had to be improved because TeX fragments misused symbols and operators e.g. f^{'} denotes derivation. A maximum depth limit was introduced into one of the algorithms which computed canonical distributivity because independent tables and other structures were put into one TeX fragment and the algorithm complexity grew rapidly with the number of operators.

4 Conclusion and Availability

There are many digital scientific repositories with little semantic information available. We think that focusing on these repositories is very important because they will still prevail in the near future. We showed that mathematical searching in one of the world's most important one is feasible at least from the technical point of view. The focus was mainly on recall so the next step is to focus on precision and preferring more similar result. This means to start using the built-in ranking algorithm and gathering feedback from the users. The online version with additional description can be found at [5]. Administration credentials, sources to Wikipedia converters, dumps and document set are available upon request.

References

1. Mišutka, J., Galamboš, L.: Extending Full Text Search Engine for Mathematical Content. In: Towards Digital Mathematics Library, Birmingham, UK, pp. 55–67 (2008)
2. Tamáš, M.: WSE (2010), http://www.projects.eblend.net/web-search-engine/
3. Wikipedia: Database download, http://www.en.wikipedia.org/wiki/Wikipedia:Database_download
4. latex2mathml converter service, http://www.tex2xml.kwarc.info
5. EgoMath2 mathematical search engine, http://www.egomath.cythres.cz:8080/egomath/