Exploiting Semantic Annotations in Math Information Retrieval

Petr Sojka
Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic
sojka@fi.muni.cz

ABSTRACT

This paper describes exploitation of semantic annotations in the design and architecture of MIaS (Math Indexer and Searcher) system for mathematics retrieval. Basing on the claim that navigational and research search are 'killer' applications for digital library such as the European Digital Mathematics Library, EuDML, we argue for an approach based on Natural Language Processing techniques as used in corpus management systems such as the Sketch Engine, that will reach web scalability and avoid inference problems. The main ideas are 1) to augment surface texts (including math formulae) with additional linked representations bearing semantic information (expanded formulae as text, canonicalized text and subformulae) for indexing, including support for indexing structural information (expressed as Content MathML or other tree structures) and 2) use semantic user preferences to order found documents.

The semantic enhancements of the MIaS system are being implemented as a math-aware search engine based on the state-of-the-art system Apache Lucene, with support for [MathML] tree indexing. Scalability issues have been checked against more than 400,000 arXiv documents.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Information Storage and Retrieval—*Content Analysis and Indexing*; I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*; I.7.3 [**Computing Methodologies**]: Document and Text Processing—*Index Generation*

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

MIaS, WebMIaS, digital mathematics libraries, information systems, math indexing and retrieval, mathematical content representation

1. INTRODUCTION

A semantic, math-aware search is a gateway to the vast treasure of knowledge in Digital [Mathematical] Libraries (DML) as EuDML [3]. There are two main types of search: *navigational* and *re*search *search*.

The goal of a navigational (exploratory) search is to locate documents or web pages related to the user's intention usually expressed as a sparse set of keywords. Research searches tend to be more fine-grained: their goal is to reveal again (re-search) evidence of a piece of previously published information—already known paper, theorem, lemma or equation. Both types of searches benefit from proper handling of *semantics*—that is, the meanings of *words* and their relationships. Both use cases benefit from possibility to narrow search by domain of interest specified by user as e.g. Mathematical Subject Classification (MSC) numbers.

To create a 'killer application' for EuDML—semantic, math-aware search for STEM digital libraries—we are developing the Math Indexer and Searcher (MIaS) [9, 10] and a user interface WebMIaS [6]. In its present form, MIaS primarily supports navigational searches, and it is unique in supporting not only words, but also mathematical formulae heavily used in STEM papers. To support better research searching and to improve navigational searching we are expanding our indexing terms with several types of *semantic annotations*—topical terms, canonical formulae terms and interlingual terms for multilingual retrieval.

2. PROBLEMS OF STATE OF THE ART

Semantic searching and the semantic web have become buzzwords today, naming different approaches to search. Google uses it for its *Knowledge Graph* of millions of interlinked words and collocations. Systems like GoPubMed build on ontology generation and usage. Wolfram Alpha believes in mathematical descriptions of a computable universe of knowledge. All are quite costly and time-consuming tasks.

Hakia, Sensebot, Powerset, DeepDyve and Cognition are further examples of 'semantic search' systems. All systems exploit context and common sense knowledge with natural language analyses of a surface form representation (bag-of-words) of documents and queries, trying to narrow the *semantic gap* between different layers of document representations: strings of optically recognized characters, words (morphology), word *n*-grams (collocations, phrases) to disambiguated word meanings and related topics which depend on an understanding of pragmatics. Narrative aspects of documents are usually neglected in current approaches, as is math formulae handling.

As clearly expressed by Jeff Dean (Google) in his Google I/O 2008 talk, the need for the scalability of web search demands a new generation of indexing design for every new order of magnitude of the number of documents. Semantic enhancements thus have to be computed, disambiguated and indexed in advance, limiting on-the-fly search query computations to linear or rather sublinear (*constant time*) algorithms. Costly *semantic inference* algorithms would increase search system response latency too much. Not enough time for inference should be compensated for by indexing precomputed multiple representations and *canonical semantic annotations* to increase the search system precision and performance (caching the index in distributed RAM is possible), as we are doing in MIaS.

ESAIR'12, November 2, 2012, Maui, Hawaii, USA.
ACM 978-1-4503-1717-7/12/11.

3. SEMANTIC CANONICAL ANNOTATIONS

We believe in an empiricist approach to the natural language processing of documents, and their retrieval enhanced by semantic canonicalized annotations. NLP-based corpora management systems like Sketch Engine [5] with underlying corpus tools [2] narrow the gap between surface text and sought after meaning. Tools permit a document to appear as a list of tokenized words in uniquely numbered positions, to add part-of-speech tags, to compute collocations and phrases using various metrics (logDice, MI-score) and to create and store additional variant semantic annotations linked to document positions.

MIaS allows the indexing of tree structures (as Presentation or Content MathML), in addition to standard 'bag-of-word' indexing, as a Lucene plug-in [9, 10], allowing scalable indexing of Digital Mathematical Libraries (DML). The processing pipe starts with documents as scanned bitmaps (almost 80% in EuDML) or born-digital PDFs. Bitmaps are processed by math-aware OCR system Infty [11], and born-digital PDFs by MaxTract program [1] to get Presentation MathML.

In addition to a canonical version of the Presentation MathML formulae tree, other normalized representations, 'annotations', are created, weighted and indexed. They represent the structure of more general (sub)formulae employing variable, constant and a common subterm unification. If unambiguous or weighted Content MathML(s) can be created from Presentation MathML, it is also indexed. Word terms of formulae expanded as vocalized for reading aloud are also indexed, as interlingual parts of a document. To minimize the number of indexed entries, the process of *canonicalization* involves converting annotation into canonical representation. In MathML tree indexing, e.g., lexicographical ordering is used for normalizing math terms with commutative operators.

We are also using the Gensim tool [8] for computing weighted document LDA topics from document representation enhanced by the above-mentioned annotations, to iteratively enhance semantic annotations by adding new topical indexing terms via multilingual LDA mappings [7], and by translated set of keywords describing paper classifications (MSC is attached to virtually all math journal paper nowadays).

Document similarities are used to weigh matched terms to compute rankings to order found documents: an improvement to match user expectation is taking into account user preferences stated as subdomains of interest specified by MSC numbers. Superdocument (concatenation) of known papers of given MSCs in DL is created and similarity of this superdocument with documents in query hits (by Gensim implementation of LDA) is used to (re)order found document list: similarity and explicit hit rankings are multiplied to get a new ranking of query hits.

4. CONCLUSIONS: TOWARDS SEMANTIC MULTILINGUAL MATH SEARCH

We have described several semantic annotations and enhancements that improve math information retrieval in DMLs. We are using the MREC corpus [6] of 438,000 preprocessed arXiv articles with 158 million mathematical formulae. for our annotation experiments and evaluation. We have found the recent paper [4] inspiring not only for suggested visual interaction with DML corpus based on topics, but also through a similar document metric that is proportional to LDA topics overlap.

Semantic annotations and techniques used during a multiple phase NLP based DML indexing workflow have proven to increase F-measure performance of WebMIaS. We hope to have the impact of semantic annotations for both navigational and research use cases evaluated by the workshop date, to demo and discuss it at ESAIR workshop.

Acknowledgements. This work has been partially supported by the the EU through its CIP ICT Policy Support Programme 'Open access to scientific information", Grant Agreement No. 250503. The author thank other MIaS and EuDML colleagues for fruitful discussions and to the paper reviewers for improvement suggestions.

5. REFERENCES

[1] Josef B. Baker, Alan P. Sexton, and Volker Sorge. MaxTract: Converting PDF to LaTeX, MathML and Text. In *AISC/DML/MKM/Calculemus*, Vol. 7362 of LNAI, pp. 422–426. Springer, 2012.

[2] Marco Baroni and Adam Kilgarriff. Large linguistically-processed web corpora for multiple languages. In *Proc. of the 11th Conference of the EACL '06*, pp. 87–90, Stroudsburg, PA, USA, 2006. ACL.

[3] José Borbinha, Thierry Bouche, Aleksander Nowiński, and Petr Sojka. Project EuDML—A First Year Demonstration. In *Proc. of 10th MKM 2011*, Vol. 6824 of LNAI, pp. 281–284, Berlin, Germany, July 2011. Springer-Verlag. `http://dx.doi.org/10.1007/978-3-642-22673-1_21`.

[4] Allison J.B. Chaney and David M. Blei. Visualizing topic models. In *Intl. AAAI Conference on Social Media and Weblogs*, Department of Computer Science, Princeton University, Princeton, NJ, USA, March 2012.

[5] Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. The Sketch Engine. In *Proc. of the 11th EURALEX International Congress*, pp. 105–116, Lorient, France, 2004.

[6] Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec. Web Interface and Collection for Mathematical Retrieval: WebMIaS and MREC. In *Proc. of DML 2011. Bertinoro, Italy, July 20–21st, 2011*, pp. 77–84. Masaryk University, July 2011. `http://hdl.handle.net/10338.dmlcz/702604`.

[7] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. Cross lingual text classification by mining multilingual topics from wikipedia. In *Proc. of the 4th ACM international conference on Web search and data mining*, WSDM '11, pp. 375–384, New York, NY, USA, 2011. ACM.

[8] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proc. of LREC 2010 workshop New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`, software available at `http://nlp.fi.muni.cz/projekty/gensim`.

[9] Petr Sojka and Martin Líška. Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In *Proc. of 10th MKM 2011*, Vol. 6824 of LNAI, pp. 228–243, Berlin, Germany, 2011. Springer-Verlag. `http://dx.doi.org/10.1007/978-3-642-22673-1_16`.

[10] Petr Sojka and Martin Líška. The Art of Mathematics Retrieval. In *Proceedings of the ACM Conference on Document Engineering, DocEng 2011*, pp. 57–60, Mountain View, CA, 2011. ACM. `http://doi.acm.org/10.1145/2034691.2034703`.

[11] Masakazu Suzuki, Fumikazu Tamari, Ryoji Fukuda, Seiichi Uchida, and Toshihiro Kanahori. INFTY — An integrated OCR system for mathematical documents. In *Proc. of ACM Symposium on Document Engineering 2003*, pp. 95–104, Grenoble, France, 2003. ACM.