Full Text Formula Search in zbMATH

Fabian Müller and Olaf Teschke (FIZ Karlsruhe, Berlin, Germany)

Three years ago, formula search has been introduced in zbMATH[1]. Formula retrieval is based on three ingredients: digitisation, content extraction, and a math-aware search engine. Our aim is to give an update on its status and developments.

The search uses the MathWebSearch engine[2] developed by the KWARC group[3] at Jacobs University Bremen, which leverages a technique called *substitution tree indexing*[4]. This method enables high-performance structural searching in a large corpus of formulae using query expressions that may contain free variables or placeholders. The latter are denoted by a leading question mark and will match arbitrary subexpressions of any complexity. When occurring multiple times in the input query, they will be substituted with the same concrete expression for each occurrence. Thus a query like "-?a \leq ?b \leq ?a" would match the formula "$-\sqrt{a} \leq f(u) \leq \sqrt{a}$", but not "$0 \leq x \leq 1$".

In order to be indexed, a LaTeX document must first be converted to MathML, which is handled by the LaTeXML converter developed by Bruce Miller at NIST[5]. After conversion, the formulae contained in each document are extracted by the indexing engine and can then be retrieved using an XML-based query syntax. The formula search interface features an interactive preview converting the user's LaTeX input on-the-fly to MathML that is then displayed in the browser.

As outlined in[1], the resulting indexes are quite huge, and processing the vast amount of zbMATH formulae requires considerable resources. Hence, it has been a non-trivial (though hidden) achievement to transform the project prototype into a sustained feature that is now updated daily along with the ever-growing inflow of documents. Likewise, the integration of new versions of LaTeXML went certainly unnoticed, though Bruce Miller's efforts have significantly improved on capturing underlying semantics supporting a more precise retrieval. However, the difficult challenge of extracting semantic content from LaTeX information still remains unsolved: E.g., searching for ?a^?n + ?b^?n=?c^?n may produce results involving Diophantine equations, a two-dimension eikonal equation, or norms fulfilling 1/p+1/q=1/r. Defining the domain of the variables would help, but this is currently not feasible; instead, the more pragmatic approach of specifying the mathematical area often serves the same purpose. Hence, it was a useful improvement to enable combined search of metadata and formulae in zbMATH two years ago, which allows for refining formula search results by terms or subjects.

The main remaining challenge to improve formula search is digitisation. Since LaTeX is required, the search was initially restricted to zbMATH reviews and abstracts that are available in this format. The situation for full texts is worse: Even articles of the last decades are usually only available as pdf (older just as scans). Though some approaches for LaTeX conversion exist[6,7], the results often lack the precision required for seamless formula indexing. Publishers could provide a tremendous support for math retrieval by making LaTeX sources or derived XML data available.

Fortunately, this is already the case for the arXiv. The recent indexing of about 120,000 arXiv full text links[8] within the zbMATH database enabled us to extend the formula search considerably. Even this small fraction of the 3.7 million zbMATH documents pushed the number of indexed formulae to more than 100 million. It is interesting to note that for a sample of frequent formulae, the number of search results increased only by an average of 30%, indicating that relevant formulae are frequently mentioned in the reviews. On the other hand, there is now a long tail of rare expressions available for searching which did not show up in the corpus before.

An interesting aspect of the zbMATH user survey conducting during this year's ECM[9] was that formula search is among the least frequently used though potentially most promising future features of zbMATH. This discrepancy is not surprising: the first reaction of most mathematicians encouraged to test formula search is that they believe such a system could not work yet. Taking the mentioned obstacles into account, there may be some justification for this; however, the progress made during the last years has surpassed our expectations, so we believe it is worth to stay up-to-date with this feature and experiment from time to time by searching for your favourite formula.

For the authors' CVs and photos we refer to the zbMATH column of the EMS Newsletter No. 99 by the same authors.

[1] M. Kohlhase et al., Eur. Math. Soc. Newsl. 89, 56–58 (2013; Zbl 1310.68217).

[2] https://github.com/KWARC/mws

[3] https://kwarc.info/

[4] P. Graf, "Substitution tree indexing", Lect. Notes Comp. Sci. 914, 117—131 (1995; doi:10.1007/3-540-59200-8_52)

[5] http://dlmf.nist.gov/LaTeXML/

[6] Infty, http://www.inftyproject.org/en/index.html

[7] Maxtract, http://www.cs.bham.ac.uk/research/groupings/reasoning/sdag/maxtract.php

[8] see F. Müller and O. Teschke, Eur. Math. Soc. Newsl. 99, 55–56 (2016; Zbl 1345.68267)

[9] A detailed report will be given in the next column.