Augmenting Mathematical Formulae for More Effective Querying & Presentation

1 Summary

Scientists and engineers search regularly for well-established mathematical concepts, expressed by mathematical formulae. Conventional search engines focus on keyword based text search today. An analogue approach does not work for mathematical formulae. Knowledge about identifiers alone is not sufficient to derive the semantics of the formula they occur in. Currently, for formula related inquiries the solution is to consult domain experts, which is slow, expensive and non-deterministic.

Consequently, core concepts to enable formula related queries on potentially large datasets are needed. While earlier attempts addressed the problem as a whole, I identify three mutually orthogonal challenges to formula search.

The first challenge, *content augmentation*, is to collect the full semantic information about individual formula from a given input. Most fundamentally, this might start with digitization of analogue mathematical content, captures the conversion from imperative typesetting instructions (i.e. TEX) to declarative layout descriptions (i.e. presentation MathML) but also deals about inferring the syntactical structure of a formula (i.e. the expression tree often represented in content MathML). In addition, this first challenge involves the association of formula metadata such as constraints, identifier definitions, related keywords or substitutions with individual formulae.

The second challenge is *content querying*. This ranges from query formulation, to query processing, actual search, hit ranking to result presentation. There are different forms of formula queries. Standard ad-hoc retrieval queries, where a user defines the information need and the math information retrieval system returns a ranked list given a particular data set. Similar is the interactive formula filter queries, where a user filters a data set interactively until she derives at the result set, which is relevant to her needs. Different are unattended queries that run in the background to assist authors during editing or readers to identify related work while viewing a certain formula.

The third challenge is *content indexing* for growing data sets. This challenge includes the scalable execution of the solutions to the two aforementioned challenges. While well-established from the area of database systems i.e. XML processing and indexing can be applied, math specific complexity problems require individual solutions.

Augmented content (challenge 1) opens up additional options for similarity search, and potentially improves the search results regardless of the applied similarity measure. In order to separate the effect of content augmentation from intrinsic improvements in the applied similarity features, I develop measures for *formula data quality* that separates those aspects. Afterwards I compare similarity measures given a certain data quality based on that quality measure. The quality measure itself is a valuable contribution. Its use is not limited to search. For example a quality measure can assist authors to check their documents for (1) missing definitions, (2) ambiguities, (3) dependency problems, and (4) redundancy. Note that I focus on quality measure for individual formulae assuming that the relevant meta-information has already been extracted from the surrounding text. Since the developed data quality measures are tailored to my approach of similarity search I'll give more details on the data quality measures after having introduced the similarity factors below.

Math search engines have used some form of similarity measure since the early days of Math search in the 2000'th. However, Youssef and Zhang [1] were the first who branded 5 factors that contribute to similarity measures in July 2014. Those factors are the starting point for my systematic approach to formula similarity measures, which extends their work in the following way:

First, I differentiate between *proper*- and *contextual* formula similarity factors. Proper factors are quantified by applying a distance measure to a formula-formula pair. I identify three proper factor groups that correspond to *lexical*, *structural* and *semantic* similarity. Typical uses of proper factors are formula clustering, auto correction and completion for formula input while editing, and hit candidate retrieval in the search context. In contrast, contextual formula similarity factors describe the relatedness between a formula *search* pattern and a candidate hit list. Even though, contextual factors are less formal and do not necessarily depend on the concrete search pattern, they play an essential role for Math Information Retrieval Applications, i.e. in search result ranking and cannot therefore be neglected.

In addition, I will describe existing MIR systems using the notation of the discussed similarity factors. This will provide a template, on using formula similarity factors as building blocks for specialized applications or future search MIR systems. Eventually, not all features can be described by the identified factors. In that case I would refine the factor list to ensure that all systems participated in the NTCIR 10 and 11 challenges can be modelled using the factors as building blocks.

Second, I analyze the impact of each individual factor and the inference patterns between different factors. Additionally, I create case studies and templates on how different factors can be combined into use-case specific similarity measures.

The identified factors imply dimensions of the aforementioned formula quality measure. Namely, I define four dimensions of formula quality: (1) typographic and lexical quality; (2) syntactic structure quality; (3) semantic quality; and (4) metadata quality. An example for low syntactic data quality is misinterpretation of $f(a + b)$ as $fa + fb$ rather than f "applied at" $(a + b)$. In a search context this might hinder relevant results from being matched. The associated quality measure for the structural data quality needs to measure to which degree the structure of the formula was captured correctly. Note that the quality measure for contextual information needs to be related to a main unit of possible relevant meta-information.

To assess the similarity measures used by current MIR systems, I will describe existing MIR systems using the uniform similarity factor approach, and perform a comparative evaluation in NTCIR Tasks. This evaluation is twofold and depends on human relevance judgement for document sections on the one hand, and on known item retrieval, on the other hand. For the impact analysis of individual similarity factors, I was using (1) gold standard driven sensitivity analysis [2] and (2) known item based evaluation [4]. The two metrics discussed above measure not only the performance of MIR systems, but they also evaluate the performance of similarity factors individually. The similarity measures evaluated were taken from existing MIR systems and additional measures proposed in the literature and taken from other disciplines. One result of this evaluation is that some factors, namely, those in the group of proper semantic similarity measures, require a minimum level of data quality to contribute to search result quality in a meaningful way.

By collecting a large and heterogeneous sample of similarity measures, I am confident to have laid a good foundation to evaluate measures that will be developed in the future. I used the existing arXiv corpus for the evaluation. In addition I created, based on the experiences gained with that corpus, an additional corpus from Wikipedia. Since the HTML formats generated in this corpus was obtained automatically from LaTeX or WikiText, respectively, data quality is not perfect, but I consider it good enough to get qualitative insights about the impact on individual measures. In addition, I use the DLMF/DRMF data-sets which are partially available in different levels of data quality to analyze the impact of data quality on formula search and individual factor effect. I expect to see that the original version of LaTeXML without any content enrichment has the lowest data quality and lowest precision in search result. The DRMF content augmentation process will have raised the data quality and also improved search results. The best results with regard to data quality and search results are expected to be obtained by using the manually generated dataset of DLMF chapter 1-4 by Zhang and Youssef.

The main contribution of this thesis is a systematic analysis of the opportunities and limitations of formula similarity search for context free formula, in dependence of the formula data quality and application scenario. Incorporation of formula unrelated information that is given in the text only, is beyond the scope of this work.