Writing Search Queries for the Math2 Task at NTCIR-11

Michael Kohlhase (Editor)
Jacobs University
http://kwarc.info/kohlhase

May 15, 2014

Abstract

This document shows how to write challenge queries for the Math2 Task at NTCIR-11.

1 Introduction

NTCIR is an evaluation workshop aimed to advance the research of Information Access technologies such as Information Retrieval, Text Summarization, Information Extraction, and Question Answering [NTC]. NTCIR-11 features the Math2 task as a successor to the very successful Math pilot task at NTCIR-10; see [AKO13] for an overview.

The Math-2 task at NTCIR-11 [NTM] aims to develop an evaluation test collection for mathematical formulae search, in order to facilitate and encourage research in mathematical information retrieval and its related fields.

Participants have received the NTCIR-Math dataset which contains ca. 8 Million files with paragraphs from 100 000 HTML full texts of articles from the Cornell Preprint arXiv [ArX] transformed with the LaTeXML system [LTX]. Formulae are marked up as MathML (presentation markup with annotated content markup and LaTeX source; see [Aus+10]).

The Math Task at NTCIR-11 is a full-text information retrieval task. Participating IR systems obtain a list of queries consisting of words and formulae (possibly) with wildcards (query variables) and return for every query an ordered list of names of file that are claimed to match the query, plus possible supporting evidence (e.g. the identifiers of formulae and the substitution for query variables). A sample of the most commonly returned hits will be judged on their relevance by human evaluators. Performance of participating systems are evaluated using a standard IR evaluation measures, precision and Mean Average Precision (MAP).

The query formats are explained [Koh14], this document shows how to write the challenge queries for the Math Task at NTCIR-11 in (augmented) LaTeX, the "working language" of Mathematics.

2 The NTCIR-11 Math2 LaTeX format

We use LaTeXML to generate the query format specified in [Koh14] from a simple LaTeX file. A topic has two kinds of information. The **public** information that is shown to the task participants (see Section 2.1) and the **private** information that is only shown to the evaluators (see Section 2.2).

Listing 1 shows the top-level structure of a LaTeX file with topics, where ⟪topics⟫ is a list of topic representations as the simple one in Listing 2. Figure 1 shows the information the participants see (where the ntcir11−topics was called with the private option as in Listing 1) and Figure 2 the extended version the evaluators and admins see for the topic in Listing 2.

Listing 1: Top-level structure of a NTCIR-11 Topics file in LaTeX

```
\documentclass{article}
\usepackage[private]{lib/ntcir11−topics}
\begin{document}
⟪topics⟫
\end{document}
```

The topic environment delineates the query (a "topic" in NTCIR parlance). The topic environment takes an argument ⟪title⟫, which gives a human-readable title to the topic. The ⟪title⟫ of a topic is is private.

2.1 Public Information for the Participants

The fquery environment encapsulates a formula schema: a LaTeX formula ⟪formula⟫[1], where query variables are specified by ?⟪name⟫. A formula schema stands for any formula in the dataset that *looks/behaves like* ⟪formula⟫, with the query variables replaced by arbitrary formulae. In the example in Listing 2 uses the formula schema a+?b, which should match formulae like $a + 1$, $a + 3n^2$, etc. Three clarifications are in order:

1. we allow subformula matches as well, e.g. a+?b would also match $(a + 3n^2) \cdot m^7$.
2. Query variables with the same name should have instances that *look/behave alike*. A formula schema ?x+?y=?x+?y would match any expression of the commutativity for $+$.
3. We do not specify what "looks/behaves like" means on purpose. The idea is that the topic expresses an information need that the query engines can satisfy in any way they see fit. The results will be evaluated by human judges on their relevance to the information need.

Listing 2: A Simple Query in LaTeX

```
\begin{topic}{Jack the Ripper adds up}
  \begin{fquery}$a+?b$\end{fquery}
  \begin{keyword}Jack\end{keyword}
  \begin{keyword}Ripper\end{keyword}
\begin{private}
    \begin{relevance}
      The hits should give an answer to the question whether there is any connection
      between Jack the Ripper and sums that start with a variable $a$.
    \end{relevance}
    \examplehit{http://example.org/files/4711/0815.xhtml}
    \contributor{Michael Kohlhase}
\end{private}
\end{topic}
```

These formula queries can be accompanied by keywords that are matched against the text. Again, the keywords only express the information need, how that is satisfied by the search engines is up to them. But the formula queries – there may be multiple ones – and the keywords should together express the information need as well as possible.

[1] macros from the AMSLaTeX packages are allowed.

NTCIR11-Math2–1: Jack the Ripper adds up

Formula Query: $a + \boxed{b}$
Keyword: Jack
Keyword: Ripper

Figure 1: The Query from Listing 2 Formatted for Participants

2.2 Private Information for the Evaluators/Organizers

The third kind of information in the formula topic is private to the evaluators and organizers; the NTCIR-11 Math2 participants do not get to see it during the challenge. Therefore it is guarded by the private environment. For LaTeX-internal reasons it is important that the **begin**{private} and **end**{private} are at the beginning of the line, i.e. do not have any preceding spaces – the LaTeX error messages are quite hard to connect to this problem.

The relevance environment encapsulates a text that gives further information to the evaluators to help them judge the relevance of results. This text – which can contain math formulae – will be shown to evaluators during the evaluation sessions.

The \examplehit macro allows to exhibit an example hit. We ask topic authors to give us at least one example hit (multiple can be specified by repeating the \examplehit macros) from the NTCIR data set. We provide two search engines for this (please contact the author).

NTCIR11-Math2–2: Jack the Ripper adds up

Formula Query: $a + \boxed{b}$
Keyword: Jack
Keyword: Ripper
Relevance: The hits should give an answer to the question whether there is any connection between Jack the Ripper and sums that start with a variable a.
Example Hit: `http://example.org/files/4711/0815.xhtml`
Contributor: Michael Kohlhase

Figure 2: The Query from Listing 2 Formatted for Judges

3 Conclusion and Availability

We have detailed how to write challenge queries for the Math2 Task at NTCIR-11. With this it becomes very easy to contribute queries to for the NTCIR-11 Math2 task. The necessary LaTeX package ntcir11−topics.sty is available at `http://kwarc.info/kohlhase/event/NTCIR11`. The tools necessary for the compilation of XML queries from LaTeX topics files are available from the author upon request.

References

[AKO13] Akiko Aizawa, Michael Kohlhase, and Iadh Ounis. "NTCIR-10 Math Pilot Task Overview". In: *NTCIR Workshop 10 Meeting*. Tokyo, Japan, 2013, pp. 1–8. URL:

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/OVERVIEW/01-NTCIR10-OV-MATH-AizawaA.pdf.

[ArX] *arxiv.org e-Print archive*. URL: http://www.arxiv.org (visited on 06/12/2012).

[Aus+10] Ron Ausbrooks et al. *Mathematical Markup Language (MathML) Version 3.0*. W3C Recommendation. World Wide Web Consortium (W3C), 2010. URL: http://www.w3.org/TR/MathML3.

[Koh14] Michael Kohlhase. *Formats for Topics and Submissions for the Math2 Task at NTCIR-11*. Tech. rep. NTCIR, 2014. URL: http://ntcir-math.nii.ac.jp/wp-content/blogs.dir/13/files/2014/05/NTCIR11-Math-topics.pdf.

[LTX] Bruce Miller. *LaTeXML: A LaTeX to XML Converter*. URL: http://dlmf.nist.gov/LaTeXML/ (visited on 03/12/2013).

[NTC] *Workshop Aims — NTCIR-11 — NTCIR*. URL: http://research.nii.ac.jp/ntcir/ntcir-11/aims.html (visited on 05/14/2014).

[NTM] *NTCIR-11 Task: Math2*. URL: http://ntcir-math.nii.ac.jp/ (visited on 02/18/2014).