Formats for Topics and Submissions for the Math2 Task at NTCIR-11

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

June 2, 2014

Abstract

This document presents the formats of the challenge queries and results for the Math2 Task at NTCIR-11. The document will be specified in further and clarified in a discussion process with the NTCIR11-Math2 participants.

1 Introduction

This document presents the challenge queries for the Math Task at NTCIR-11 [NTM]. 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). Results will be evaluated using a standard IR evaluation measures, precision and Mean Average Precision (MAP).

2 Query Formats

The general form of queries is given in Figure 1 (see appendix A.1 for a RelaxNG schema). An example is given in Appendix A.2.

Each topic has an identifier of the form NTCIR11-Math2-⟪num⟫ in the num element. The query itself is given in the query element and consists of a list of words and formula schemata given the keyword and formula elements. The contents of the keyword is a UniCode string in UTF-8 encoding which is interpreted as a single search keyword. The keyword and formula elements carry an id element with an identifier that can be used for identification in the results (see Section 3).

The content of the formula element is a MathML formula in parallel markup (see section 5.4 in [Aus+10]). This contains three representations of the formula schema: ⟪CMML⟫ in content MathML, ⟪PMML⟫ in presentation MathML, and ⟪LaTeX⟫ as the LaTeXsource. Note that the

```
<?xml version="1.0" encoding="utf-8"?>
<topics xmlns="http://ntcir-math.nii.ac.jp/"
        xmlns:m="http://www.w3.org/1998/Math/MathML">
  <topic>
    <num>NTCIR11-Math2-⟪num⟫</num>
    <query>
      <keyword id="⟪id⟫">⟪keyword 1⟫</keyword>
      ...
      <keyword id="⟪id⟫">⟪keyword n⟫</keyword>
      <formula id="⟪id⟫">
        <m:math>
          <m:semantics>
            ⟪CMML⟫
            <m:annotation-xml encoding="MathML-Presentation">
              ⟪PMML⟫
            </m:annotation-xml>
            <m:annotation encoding="application/x-tex">⟪LaTeX⟫</m:annotation>
          </m:semantics>
        </m:math>
      </formula>
      ...
      <formula id="⟪id⟫">...</formula>
    </query>
  </topic>
  <topic>...</topic>
</topics>
```

Figure 1: Machine-Readable form of Queries

formula schema can contain query variables represented as qvar elements (in MathML) and \qvar macros in LaTeX both specify the name of the variable, which can be reference in the result justification.

Note furthermore that the conversion to content MathML is heuristic and not always optimal. But we use the same conversion in the dataset and queries, so errors cancel out in practice. In any case, the <m:annotation> child of the MathML expression has the original TeX formula, so alternative conversions can be employed.

3 Reporting Results

Participants report the results of up to four "runs" of their search engine with the given queries (see Section 2) over the dataset supplied. A "run" is your system outcome for a given approach. In general, you can imagine a run as the outcome of a specific variant of your system testing a specific hypothesis. Each run contains exactly 1000 search results. Note that even if the search method only finds fewer results, then the 1000 have to be filled up – e.g. by random selection.

Results of these runs are reported by e-mailing a results file to `ntcir11adm-math@nii.ac.jp`. Please name your file as ⟪group-id⟫.⟪ext⟫ and decorate every result with a ⟪runtag⟫ of the form ⟪group-id⟫-⟪run-id⟫, where ⟪group-id⟫ is the group identifier you have chosen

upon NTCIR-11 registration and ⟪run-id⟫ can be chosen freely. Results can be reported in two forms depending whether they have justifications.

3.1 Simple Results

Results to Math queries can be reported as lists of six-tuples (one line per hit)

NTCIR11−Math−⟪num⟫ 1 ⟪filename⟫ ⟪rank⟫ ⟪score⟫ ⟪runtag⟫

where

1. The first column contains the query identifier.
2. The second one is fixed to 1.
3. ⟪filename⟫ the full path of the file that purportedly matches that query in the NTCIR11-Math2 dataset, e.g. `xhtml/1/hep-th0007174/hep-th0007174_1_1.xhtml`.
4. ⟪rank⟫ gives the rank of the "hit" in the answers to the particular query. ⟪rank⟫ should start from 1 and should be always incremental (by one, i.e. the rank numbers just represent the number of lines; no documents have the same rank).
5. ⟪score⟫ specifies how "happy" your system is with the result. Make sure that ⟪score⟫ and ⟪rank⟫ are consistent; for example, $\langle\!\langle \text{rank} \rangle\!\rangle_i < \langle\!\langle \text{rank} \rangle\!\rangle_j$ iff $\langle\!\langle \text{score} \rangle\!\rangle_i > \langle\!\langle \text{score} \rangle\!\rangle_j$.
6. Finally, ⟪runtag⟫ identifies the "system run" (see above).

3.2 Results with Justifications

Results with justifications should be reported in an XML file structured as in Figure 2 (see Appendix A.3 for a RelaxNG schema).

```
<?xml version="1.0" encoding="utf−8"?>
<results xmlns="http://ntcir−math.nii.ac.jp/">
  <run runtag="⟪runtag⟫" runtime="⟪ms⟫" run_type="⟪type⟫">
    <result for="NTCIR11−Math−⟪num⟫" runtime="⟪ms⟫">
      <hit id="⟪id⟫" xref="⟪filename⟫" score="⟪score⟫" rank="⟪rank⟫"/>
      ...
      <hit .../>
    </result>
    ...
    <result>...</result>
  </run>
  ...
  <run>...</run>
</results>
```

Figure 2: Returning Results in XML

For each run there is a run element whose runtag attribute specifies the run identifier as discussed above. Runs can be manual or automatic, correspondingly the value of ⟪type⟫ is one of automatic or manual. For each query there is a result element, whose for attribute identifies the query. The total run time in milliseconds used to answer the query – including query processing and result generation – is specified in the runtime attribute; its value should be an integer. This attribute is mandatory for automatic runs.

⟪filename⟫, ⟪rank⟫, ⟪score⟫, and ⟪runtag⟫ are subject to the constraints specified in Section 3.1.

The result element contains a list of hit elements that identify a file from the NTCIR-11 dataset purported to match that query via the its name ⟪filename⟫ in the xref attribute. ⟪rank⟫ and ⟪score⟫ are as above and underlie the same constraints.

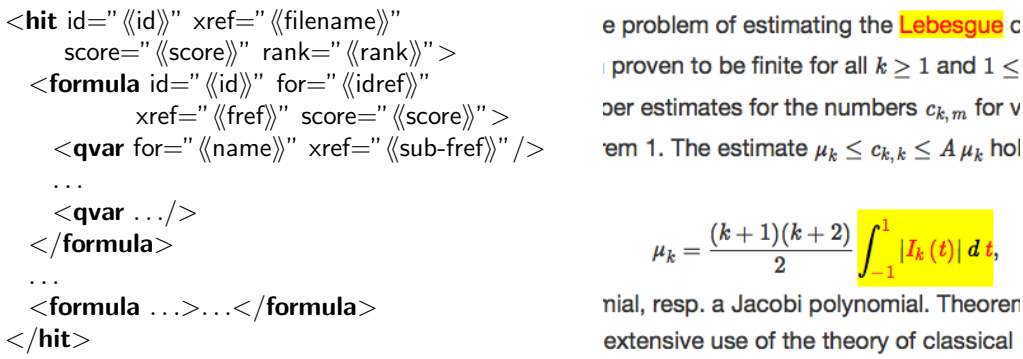

Figure 3: Justifying and Highlighting Hits

A hit can be (optionally[1]) further specified by formula elements in the hit element. Their for attributes identify the respective formula schema from the query by referencing its id attribute and specify the exact occurrences in the file via their xref attributes. The value of the xref attribute is a URI reference of the form ⟪filename⟫#⟪id⟫, where ⟪filename⟫ is the file in the dataset with the hit and ⟪idref⟫ the value id attribute of the formula in the source (for formula). Individual formula occurrences can be given a ⟪score⟫ as well in the optional score attribute.

formula elements can be further justified by giving a substitution for the query variables: For each wildcard (query variable mws:qvar whose name attribute has value ⟪name⟫) in the query, the subformula it matches can be given in a qvar element. Its for attribute specifies the ⟪name⟫ and its xref attribute points to the id of the subformula.

References

[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`.

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

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

[1]Note that the justifications will not be judged per se, but will make life of the evaluators easier and allow them to understand why the hit is justified.

A Appendix

We provide the RelaxNG schemata and examples for the XML query and result formats for convenience, they can be found `https://svn.mathweb.org/repos/NTCIR-Math/topics/lib`.

A.1 RelaxNG Schema for NTCIR Queries

We first have a schema for the participants

```
# A RelaxNG for NTCIR−11 Math Task Topics (common part)
# Id : NTCIR11 − topic − participants.rnc1282014 − 05 − 1908 : 51 : 23Zkohlhase
# HeadURL : https : //svn.mathweb.org/repos/NTCIR − Math/topics/ntcir11/lib/NTCIR11 − topic − participants.r
# (c) 2014 Michael Kohlhase, released under the GNU Public License (GPL)

namespace mws = "http://search.mathweb.org/ns"
namespace m = "http://www.w3.org/1998/Math/MathML"
namespace xsd = "http://www.w3.org/2001/XMLSchema−datatypes"
default namespace = "http://ntcir−math.nii.ac.jp/"

start = element topics {topic∗}

id.att = attribute id {xsd:ID}
formula = element formula {id.att, math}
keyword = element keyword {id.att, inline.model}

qvar = element mws:qvar {attribute name {text}}

math = grammar {include "../../lib/mathml3/mathml3.rnc"
                        PresentationExpression |= parent qvar
                        ContExp |= parent qvar}
inline.class = math
inline.model = (text | inline.class)∗

query = element query {formula∗ & keyword∗}
num = element num {text}

topic.model = num & query
topic = element topic {topic.model}
```

And then we extend it for the judges with the private parts.

```
# A RelaxNG for NTCIR−11 Math Task Topics
# Id : NTCIR11 − topic − judges.rnc1282014 − 05 − 1908 : 51 : 23Zkohlhase
# HeadURL : https : //svn.mathweb.org/repos/NTCIR − Math/topics/ntcir11/lib/NTCIR11 − topic − judges.rnc
# (c) 2014 Michael Kohlhase, released under the GNU Public License (GPL)

namespace mws = "http://search.mathweb.org/ns"
namespace m = "http://www.w3.org/1998/Math/MathML"
namespace xsd = "http://www.w3.org/2001/XMLSchema−datatypes"
default namespace = "http://ntcir−math.nii.ac.jp/"

include "NTCIR11−topic−participants.rnc"

private = element private {examplehit* & contributor* & title? & note* & relevance? & reference? }
examplehit = element examplehit {attribute href {xsd:anyURI}}
contributor = element contributor {inline.model}
title = element title {inline.model}
relevance = element relevance {inline.model}
reference = element reference {xsd:anyURI}
note = element note {inline.model}

topic.model &= private
```

A.2 An Example Query

We first have a an example of what the participants see

```
<?xml version="1.0" encoding="UTF-8"?>
<topics xmlns="http://ntcir-math.nii.ac.jp/" xmlns:m="http://www.w3.org/1998/Math/MathML">
  <topic>
    <num>NTCIR11-Math-1</num>
    <query>
      <formula id="f1.0">
        <m:math>
          <m:semantics xml:id="m1.1a" xref="m1.1.pmml">
            <m:apply xml:id="m1.1.4" xref="m1.1.4.pmml">
              <m:plus xml:id="m1.1.2" xref="m1.1.2.pmml"/>
              <m:ci xml:id="m1.1.1" xref="m1.1.1.pmml">a</m:ci>
              <mws:qvar xmlns:mws="http://search.mathweb.org/ns" name="b"/>
            </m:apply>
            <m:annotation-xml encoding="MathML-Presentation" xml:id="m1.1.pmml" xref="m1.1">
              <m:mrow xml:id="m1.1.4.pmml" xref="m1.1.4">
                <m:mi xml:id="m1.1.1.pmml" xref="m1.1.1">a</m:mi>
                <m:mo xml:id="m1.1.2.pmml" xref="m1.1.2">+</m:mo>
                <mws:qvar xmlns:mws="http://search.mathweb.org/ns" name="b"/>
              </m:mrow>
            </m:annotation-xml>
            <m:annotation encoding="application/x-tex" xml:id="m1.1b" xref="m1.1.pmml">a+\qvar{b}</m:annotation>
          </m:semantics>
        </m:math>
      </formula>
      <keyword id="w1.0">Jack</keyword>
      <keyword id="w1.1">Ripper</keyword>
    </query>

  </topic>
</topics>
```

and then one of what the judges see.

```
<?xml version="1.0" encoding="UTF-8"?>
<topics xmlns="http://ntcir-math.nii.ac.jp/" xmlns:m="http://www.w3.org/1998/Math/MathML">
  <topic>
    <num>NTCIR11-Math-1</num>
    <query>
      <formula id="f1.0">
        <m:math>
          <m:semantics xml:id="m1.1a" xref="m1.1.pmml">
            <m:apply xml:id="m1.1.4" xref="m1.1.4.pmml">
              <m:plus xml:id="m1.1.2" xref="m1.1.2.pmml"/>
              <m:ci xml:id="m1.1.1" xref="m1.1.1.pmml">a</m:ci>
              <mws:qvar xmlns:mws="http://search.mathweb.org/ns" name="b"/>
            </m:apply>
            <m:annotation-xml encoding="MathML-Presentation" xml:id="m1.1.pmml" xref="m1.1">
              <m:mrow xml:id="m1.1.4.pmml" xref="m1.1.4">
                <m:mi xml:id="m1.1.1.pmml" xref="m1.1.1">a</m:mi>
                <m:mo xml:id="m1.1.2.pmml" xref="m1.1.2">+</m:mo>
                <mws:qvar xmlns:mws="http://search.mathweb.org/ns" name="b"/>
              </m:mrow>
            </m:annotation-xml>
            <m:annotation encoding="application/x-tex" xml:id="m1.1b" xref="m1.1.pmml">a+\qvar{b}</m:annotation>
          </m:semantics>
```

```
        </m:math>
      </formula>
      <keyword id="w1.0">Jack</keyword>
      <keyword id="w1.1">Ripper</keyword>
    </query>
    <private>
      <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 <m:math><m:semantics xml:id="m2.1a" xref="m2.1.pmml"><m:ci xml:i
</relevance>
      <examplehit href="http://example.org/files/4711/0815.xhtml"/>
      <contributor>Michael Kohlhase</contributor>
    </private>
  </topic>
</topics>
```

A.3 RelaxNG Schema for NTCIR Results

```
# A RelaxNG for NTCIR-11 Math Task Results
# $Id: NTCIR11-results.rnc 1362 2014-05-23 08:55:39Z kohlhase $
# $HeadURL: https://svn.mathweb.org/repos/NTCIR-Math/topics/ntcir11/lib/NTCIR11-results.rnc $
# (c) 2014 Michael Kohlhase, released under the GNU Public License (GPL)

namespace xsd = "http://www.w3.org/2001/XMLSchema-datatypes"
default namespace = "http://ntcir-math.nii.ac.jp/"

start = element results {run+}

id.att = attribute id {xsd:ID}
runtag.att = attribute runtag {text}
for.att = attribute for {text}
xref.att = attribute xref {xsd:anyURI}
score.att = attribute score {xsd:decimal}
runtime.att = attribute runtime {xsd:nonNegativeInteger}
manualrun.att = attribute run_type {"manual"}
autorun.att = attribute run_type {"automatic"}
run.atts = (autorun.att & runtime.att) | (manualrun.att & runtime.att?)
rank.att = attribute rank {xsd:positiveInteger}

run = element run {runtag.att & run.atts & result+}
result = element result {id.att & for.att & runtime.att & hit+}
hit = element hit {id.att & score.att & xref.att & rank.att& formula*}
qvar = element qvar {for.att & xref.att}
formula = element formula {id.att & score.att & xref.att & for.att & qvar*}
```

A.4 An Example of Justified Results

```
<?xml version="1.0" encoding="utf-8"?>
<results xmlns="http://ntcir-math.nii.ac.jp/">
  <run runtag="ex1" runtime="34567" run_type="automatic">
    <result id="ex1.1" for="NTCIR11-Math-3" runtime="245">
      <hit id="foo" xref="filename" score=".7777" rank="1">
        <formula id="foof" for="idref" xref="#fref" score=".555">
          <qvar for="name" xref="#fref.0.0.7.3"/>
        </formula>
      </hit>
    </result>
    </run>
</results>
```