KWARC Blue Note*
Datsets and Mathematical Queries for the MIR/NTCIR-Math Workshops

Michael Kohlhase
Computer Science, Jacobs University Bremen
http://kwarc.info/kohlhase

August 18, 2011

Abstract

This note sketches some queries for the MIR and NTCIR-Math Workshops.

Contents

*Inspired by the "blue book" in Alan Bundy's group at the University of Edinburgh, KWARC blue notes, are documents used for fixing and discussing ϵ-baked ideas in projects by the KWARC group (see http://kwarc.info). Unless specified otherwise, they are for project-internal discussions only. Please only distribute outside the KWARC group after consultation with the author.

1 Introduction

The MIR 2012 workshop at CICM [Cic] in Bremen will feature a friendly competition for the systems presented at the workshop: the Math IR Happening [Mir]. This event will also serve as a dry run for the NTCIR Math Track, a pilot task in the NTCIR-10 Workshop for Evaluation of Information Access Technologies [Ntc].

1.1 The Math IR Happening

Since math information retrieval is still quite young and developing, we will not make this an official competition, but a happening, where we get together and test our system on a common set of problems. We expect the happening to transcend the workshop proper.

The aim of the MIR happening is to jointly gain a better understanding into the information retrieval needs of mathematicians and the respective strengths and weaknesses of the respective IR approaches and systems. As a tangible result of the happening the organizers will compile a survey paper and report of this newly-gained understanding.

In particular, it is not an aim of the MIR happening to determine "winners" of the competition in any form. That may be an aim of a subsequent competition, when we have a better grip on the problems and possible evaluation approaches. MIR Challenges

1.2 The NTCIR Workshop

The NTCIR workshop will take place in Spring 2013, even though it is still a pilot task (and thus not fully part of NTCIR-10), it will be and will be more formal

This note sketches example presents a practice data set for participants and example queries.

1.3 The MIR/NTCIR-Math Challenges

We plan to conduct the happening via three challenges (details to be worked out further):

Formula Search (Automated) in the categories:

- similarity search for formulae
- instance search (query formulae with query variables)

The judges select/prepare a formula database and a set of formula queries. The formula database contains a list of formulae with identifiers. Every formula in two encodings: LaTeX and MathML (parallel-markup presentation/content). The query formulae are in the same format (extended by query variables). Competing IR systems obtain the formula database and the list of formula queries and return for every query an ordered list of "hits" (identifiers of formulae claimed to match the query), plus possible supporting evidence (e.g. a substitution for instance queries). Results will be judged on precision, recall, results ordering, and search time.

Full-Text Search (Automated) This is like formula search above, only that that we use a document collection (LaTeX and XHTML+MathML(parallel)) and a set of text/formula queries (in the same formats) instead of pure formulae. IR results are ordered lists of "hits" (i.e. XPointer references into the documents with a highlighted result fragments plus supporting evidence) and will be judged on precision, recall, results ordering, search time, and presentation of the "hits".

Open Information Retrieval (Semi-Automated) In contrast to the first two challenges, where the systems are run in batch-mode (i.e. without human intervention), in this one mathematicians will challenge the (human) contestants to find specific information in a document corpus via human-readable descriptions (natural language text), which are translated by the

contestants to their IR systems. Results to be delivered are "hits" in free form together with a description of how the results were found.

We will invite a panel of mathematicians participating in CICM as a "panel of judges" who will select/prepare the MIR challenges, judge the solutions of the contestants, and provide overall feedback.

1.4 MIR/NTCIR-Math Data set

We have prepared a practice data set for contestants, we have selected 10000 documents from the Cornell Preprint arXiv [Arx] transformed to XHTML+MathML with the LaTeXML converter [Sta+10] It can be downloaded as `http://arxmliv.kwarc.info/ntcir-10/mir-sandbox.tar.gz` (293 MB). This set of documents is the basis for the "Full-Text Search" and "Open IR" subtask sketched above. For the "Formula Retrieval" subtask we provide an excerpt in the form of MWS harvest (each harvest is essentially a list of 10000 formulae (presentation/-content MathML in parallel markup; see [Aus+10, Chapter 5]); see [KP] for details) at `http://arxmliv.kwarc.info/ntcir-10/mir-harvests.tar.gz` (155 MB; 1.6 million formulae).

Note that this data-set is only intended for preparation for the Math IR happening and NTCIR challenge, and may not be distributed. In particular the download URLs are password-protected, please contact the organizers for access.

For the MIR happening itself, we will supply a different dataset of comparable size, and for the NTCIR Challenge a data set of tenfold size. In the following we will give some example queries.

Acknowledgements

The author is indebted to Deyan Ginev for preparing the data set, to Herbert Jaeger and a colleague of Akio Aizawa for preparing example queries.

2 Formula Search

The query formulae are in the same format (extended by query variables). Competing IR systems obtain the formula database and the list of formula queries and return for every query an ordered list of "hits" (identifiers of formulae claimed to match the query), plus possible supporting evidence (e.g. a substitution for instance queries).

As similarity queries and instance queries are very similar, we will handle them at the same time, deriving them from the same example.

2.1 A Simple Arithmetic Expression

In the first example we search for a simple arithmetic expression in $\frac{1}{2}\left(\frac{p-2}{p-1}\right)^{p-1}$ (represented as `\frac12\left(\frac{p-2}{p-1}\right)^{p-1}` in TeX/LaTeX). In presentation MathML, this is represented as

Listing 1: A simple arithmetic expression in Presentation MathML

```
<mrow>
  <mfrac><mn>1</mn><mn>2</mn></mfrac>
  <mo>&#x2062;</mo>
  <msup>
    <mfenced>
      <mfrac>
        <mrow><mi>p</mi><mo>−</mo><mn>2</mn></mrow>
        <mrow><mi>p</mi><mo>−</mo><mn>1</mn></mrow>
      </mfrac>
    </mfenced>
    <mrow><mi>p</mi><mo>−</mo><mn>1</mn></mrow>
  </msup>
```

</mrow>

And this expression can be directly used as a similarity search query, which should yield a hit in document f000056.xhtml, where equation (5) has a subterm that is identical. The content MathML representation of this can directly be used as an instance query for instantiation searh (without query variables.) and should give the same hit.

Listing 2: A simple arithmetic expression in Content MathML

```
<apply>
  <times/>
  <apply><divide/><cn>1</cn><cn>2</cn></apply>
  <apply>
    <exp/>
    <apply>
      <divide/>
      <apply><minus/><ci>p</ci><cn>2</cn></apply>
      <apply><minus/><ci>p</ci><cn>1</cn></apply>
    </apply>
    <apply><minus/><ci>p</ci><cn>1</cn></apply>
  </apply>
</apply>
```

Note that we can derive other queries (more difficult) from the ones above, e.g. by renaming variable names in Listing 1 to test similarity search.

2.2 Commutativity

The next type of queries consists of MathML expressions with query variables in them. For instance in the queries in Figure 1 we are instances the commutativity law. In the similarity query on the left of the figure, we can either trust the similarity measure to allow variations on variables, or introduce wildcards (if querying supports them). In the instance query on the right side of Figure 1 we can use query variables in the mws:qvar elements directly; in the formulae, we indicate them by boxes with letter in them for better readability. Note that in this case, instance search surpasses regular expression search, since it can insist on the two "wildcards" being identical.

Presentation	Content
$op(x,y) = op(y,x)$	$\boxed{op}(\boxed{x},\boxed{y}) = \boxed{op}(\boxed{y},\boxed{x})$
<mrow> <mrow> <mi>op</mi> <mfenced open="(" close=")"> <mrow><mi>x</mi><mo>,</mo><mi>y</mi></mrow> </mfenced> </mrow> <mo>=</mo> <mrow> <mi>op</mi> <mfenced open="(" close=")"> <mrow><mi>y</mi><mo>,</mo><mi>x</mi></mrow> </mfenced> </mrow> </mrow>	<apply> <eq/> <apply> <mws:qvar>op</mws:qvar> <mws:qvar>x</mws:qvar> <mws:qvar>y</mws:qvar> </apply> <apply> <mws:qvar>op</mws:qvar> <mws:qvar>y</mws:qvar> <mws:qvar>x</mws:qvar> </apply> </apply>

Figure 1: Commutativity

Note that in the presentation case, we should also test variant queries which use

<mo fence="true">(</mo>...</mo fence="true">

instead of

<mfenced open="(" close=")''>...</mfenced>

to test for normalization. Similarly, on the content search side, we should test for the syntactic variant that uses <csymbol cd="relation1">eq</csymbol> instead of <eq/>.

2.3 Left-Hand Side of Hölder's Inequality

The next query introduces another difficulty: bound variables (here in integrals), which can be freely renamed without changing (mathematical) meaning. In the challenges, we will have queries that may or may not have the "right" variable names. The query in Figure 2 has the left hand side of Hölder's famous inequality; we search for it in situations e.g. where we want to approximate $\int_{\mathbb{R}^2} |\sin(t)\cos(t)| dt$ from above, and have generated the query by manually variablizing the specifics (here the domain $\mathbb{R}^2$ and functions sin and cos).

Presentation	Content
$\int_D \lvert f(x)g(x)\rvert\, dx$	$\int_{\boxed{D}} \lvert \boxed{f}(\boxed{x})\boxed{g}(\boxed{x})\rvert\, d\boxed{x} \leq \boxed{R}$
<mrow> <msub> <mo>∫</mo> <mi>D</mi> </msub> <mfenced open="|" close="|"> <mrow> <mi>f</mi> <mfenced open="(" close=")"> <mi>x</mi> </mfenced> <mo>&invisibletimes;</mo> <mi>g</mi> <mfenced open="(" close=")"> <mi>x</mi> </mfenced> </mrow> </mfenced> <mi>d</mi> <mi>x</mi> </mrow>	<apply> <leq/> <bind> <apply> <int/> <mws:qvar>D</mws:qvar> </apply> <bvar><mws:qvar>x</mws:qvar></bvar> <apply> <apply> <abs/> <apply> <times/> <apply> <mws:qvar>f</mws:qvar> <mws:qvar>x</mws:qvar> </apply> <apply> <mws:qvar>g</mws:qvar> <mws:qvar>x</mws:qvar> </apply> </apply> </apply> </apply> </bind> <mws:qvar>R</mws:qvar> </apply>

Figure 2: Left-Hand Side of Hölder's Inequality

3 Full-Text Search

This is like formula search above, only that that we use a document collection (LaTeX and XHTML+MathML(parallel)) and a set of text/formula queries (in the same formats) instead of pure formulae. IR results are ordered lists of "hits" (i.e. XPointer references into the documents with a highlighted result fragments plus supporting evidence) and will be judged on precision, recall, results ordering, search time, and presentation of the "hits".

3.1 Word-Formula Queries

The question

> Is there any example of a Campedelli surface with fundamental group $\mathbb{Z}_2^{\oplus 3}$?

can be expressed as a word/formula query with the words "Campedelli surface" together with the formula $\boxed{\mathbb{Z}_2^{\oplus 3}}$[1]. A known positive for this is `http://arxiv.org/abs/1205.2439` whose abstract contains

> Moveover, we construct a new Campedelli surface with fundamental group $\mathbb{Z}_2^{\oplus 3}$...

Note that this query is useful in practice as only a few researchers know the concept of a "Campedelli surface".

4 Open Information Retrieval

In contrast to the first two challenges, where the systems are run in batch-mode (i.e. without human intervention), in this one mathematicians will challenge the (human) contestants to find specific information in a document corpus via human-readable descriptions (natural language text), which are translated by the contestants to their IR systems. Results to be delivered are "hits" in free form together with a description of how the results were found.

4.1 W-Questions

The simplest kind of open information retrieval questions are W-Questions, like the following one:

> What inequalities are satisfied by Chern classes of algebraic surfaces?

Here one expected answer ist "Miyaoka-Yau Inequality"; the information can be retrieved from `http://arxiv.org/abs/0812.0462`, Which contains the paragraph

> If M is a projective n-manifold with ample canonical bundle $\mathcal{K}_M$, there exists a Kähler-Einstein metric ω with negative scalar curvature by Yau's theorem on the Calabi conjecture ([Ya2]), which was obtained by Aubin independently ([Aubin]). As a consequence, there is an inequality for Chern numbers, the Miyaoka-Yau inequality,
>
> $$(\frac{2(n+1)}{n}c_2(M) - c_1^2(M)) \cdot (-c_1(M))^{n-2} \geq 0, \tag{1}$$
>
> where $c_1(M)$ and $c_2(M)$ are the first and the second Chern classes of M (c.f. [Ya1]).

In this context, "projective n-manifold" serves as a superordinate concept to "algebraic surface" (although not always so from a mathematician point of view).

4.2 Questions for non-content

An interesting situation for math information retrieval is induced by the query

> What's the formula for Seshadri constants of $K3$-surfaces?

as no generalized known formula exists for Seshadri constants of $K3$-surfaces. Instead, the paper at `http://arxiv.org/abs/1205.2982` enumerates all that is known about Seshadri constants of $K3$-surfaces, and shown theorems for special cases.

[1] We will denote the MathML representation of a formula A with $\boxed{A}$

4.3 Existential Queries

We can directly pose the question from example 3.1 and let the contestants come up with the word/formula query presented there. But we can also make the question more difficult by asking

> Is there any example of an algebraic surface of general type with $p_g = q = 0$ and whose fundamental group is $\mathbb{Z}_2^{\oplus 3}$?

thereby avoiding the direct term Campedelli surface, which is given in terms of p_g and q witch are are often used to represent invariants for algebraic surface.

4.4 A Researcher's Dream

The following question complex was given to us by Herbert Jaeger from Jacobs University when asked what he would really like to ask a math information retrieval system. It seems to me that an answer to this is far in the future, but this is what we should try for.

> I am looking for papers where the nature of top-down processing paths in robot control architectures is discussed. "Robot control architecture" is here a catchall term; in fact I am generally interested in autonomous agent architectures, whether biological, cognitive, or robotic (where the latter is the ultimate target I am currently concerned with). Background: there is much confusion as to the nature of top-down processing paths. This confusion is both conceptual and – in simulation models – computational. Signals that are passed top-down are variously interpreted as Bayesian priors (today's top fashion), expectations, goals, values and rewards, control targets, modulation signals, coordinate transforms, and more. A main difficulty that a researcher is confronted with when entering this jungle is the heterogeneity of research contexts, epistemological biases, formal methods, biological vs. technological objectives etc. I want to get a clear picture of what are the current "main contenders", or main "conceptual principal components" in this high-dimensional conceptual and modeling space. I would be happy about comprehensive discussion papers which do not want to sell their author's pet perspective in the first place. The more formal and "mathy" the better. That is, I do not (only) want a conceptual (or even philosophical) discussion, but a clarification of the dynamical/computational behavior which is intended by / implied by the various alternatives. Important, too: how learning processes on various timescales are connected to to top-down processing; and how long-term dynamical stability of learning systems with top-down processing can be understood or guaranteed.

5 Conclusion

We have presented some example queries for the MIR and NTCIR-Math challenges. We will keep updating this blue note with additional queries, as time goes on.

References

[Arx] *arxiv.org e-Print archive*. URL: `http://www.arxiv.org` (visited on 01/08/2010).

[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`.

[Cic] *Conference on Intelligent Computer Mathematics (CICM)*. URL: `http://cicm-conference.org` (visited on 03/18/2012).

[KP] Michael Kohlhase and Corneliu Prodescu. *MathWebSearch Manual*. Web Manual. Jacobs University. URL: `https://svn.mathweb.org/repos/mws/doc/manual/manual.pdf` (visited on 04/07/2012).

[Mir] *Math IR Happening at MIR 2012*. URL: http://cicm2012.cicm-conference.org/cicm.php?event=mir&menu=happening (visited on 03/18/2012).

[Ntc] *NTCIR Math Track – Pilot Task*. URL: http://ntcir-math.nii.ac.jp/ (visited on 05/18/2012).

[Sta+10] Heinrich Stamerjohanns et al. "Transforming large collections of scientific publications to XML". In: *Mathematics in Computer Science* 3.3 (2010): *Special Issue on Authoring, Digitalization and Management of Mathematical Knowledge*. Ed. by Serge Autexier, Petr Sojka, and Masakazu Suzuki, pp. 299–307. URL: http://kwarc.info/kohlhase/papers/mcs10.pdf.