INSTITUTE OF
COMPUTER SCIENCE
Masaryk University

Mathematics Index and Search in DSpace

Vlastimil Krejčíř, `krejcir@ics.muni.cz`

OAI 10 DSpace User Group Meeting, Geneva, 20th June 2017

Motivation

In 2005, we started to build (upon DSpace + XMLUI) the **Czech Digital Mathematics Library (DML-CZ)**. Later on, the question *"What about a mathematics formulae search?"* appeared.

- Simple search based on text keywords is not sufficient for mathematical content because
- representation of thoughts in math = formulae, so
- math formulae search brings a great benefit to mathematicians (and related science disciplines).

Lets look what we struggle with...

Simple Text Search

- Plain text indexing and searching is 'easy'
- Known tools and techniques
- Searching for DSpace ⇒ the query is *'dspace'*
  - or any substring of this query...
- Matching is done character by character

...but for math formulae the situation is much more complicated!

Mathematics Formulae Index and Search

Questions/thoughts:

- How to represent a formula?
- How to index a formula?
- How to write a query?
- How to match a query and weight results?

Consider that:

- symbols and graphics heavily used
- big complexity of formulae
  - subformulae usually have sense
- variables, constants, ...
- syntax ambiguity
  - from the 'search and index' point of view

Example: Ambiguity

These make the things really hard:

$$0.5 = \frac{1}{2} = 2^{-1}$$

$$\sqrt{8} = 2\sqrt{2}$$

It is easy to find a lot of similar examples...

Example 2: Pythagorean Theorem

Pythagorean Theorem

$$a^2 + b^2 = c^2$$

Example 2: Pythagorean Theorem

Pythagorean Theorem

$$a^2 + b^2 = c^2$$

is equivalent to

$$b^2 + a^2 = c^2$$

Example 2: Pythagorean Theorem

Pythagorean Theorem

$$a^2 + b^2 = c^2$$

is equivalent to

$$b^2 + a^2 = c^2$$

is equivalent to

$$x^2 + y^2 = z^2$$

Example 2: Pythagorean Theorem

Pythagorean Theorem

$$a^2 + b^2 = c^2$$

is equivalent to

$$b^2 + a^2 = c^2$$

is equivalent to

$$x^2 + y^2 = z^2$$

and is special case of **Fermat's Last Theorem**

$$a^n + b^n = c^n$$

Besides, see book: Simon Singh: Fermat's Last Theorem

MathML

MathML - XML formulae representation

```
<math>
  <mfrac>
    <mn>1</mn>
    <msup>
      <mi mathvariant="bold">x</mi>
      <mn>2</mn>
    </msup>
  </mfrac>
</math>
```

$$\frac{1}{x^2}$$

How to get MathML

To get MathML out of existing articles – very hard task...

...our real pain :-(.

There are some tools:

- InftyReader (OCR tool)
- LaTexML
- MATLAB
- 'hand made'
- ...

MIaS: Mathematics Index and Search tool

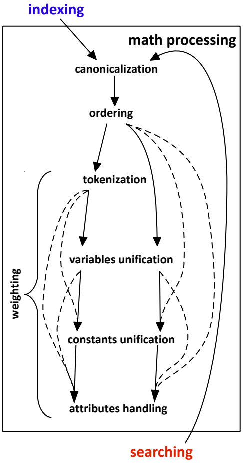
MIaS is the Java tool that provides the necessary job:
MathML ⇒ Lucene:

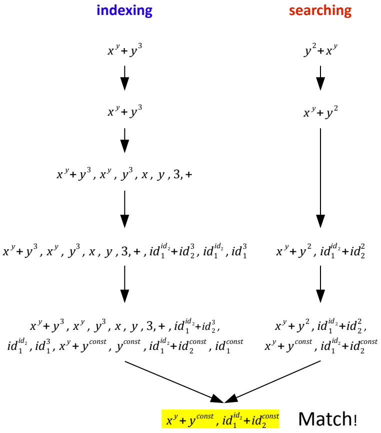
- canonicalization
- ordering
- tokenization
- variables and constant unification

The result is *M-term*, MIaS processed and plain text coded formula:

F(N(1)J(I[V=B](1)N(2)))

MIaS: schema


indexing
math processing
canonicalization
ordering
tokenization
weighting
variables unification
constants unification
attributes handling
searching


indexing
searching
$x^y + y^3$
$y^2 + x^y$
$x^y + y^3$
$x^y + y^2$
$x^y + y^3, x^y, y^3, x, y, 3, +$
$x^y + y^3, x^y, y^3, x, y, 3, +, id_1^{id_2} + id_2^3, id_1^{id_2}, id_1^3$
$x^y + y^2, id_1^{id_2} + id_2^2$
$x^y + y^3, x^y, y^3, x, y, 3, +, id_1^{id_2} + id_2^3,$
$id_1^{id_2}, id_1^3, x^y + y^{const}, y^{const}, id_1^{id_2} + id_2^{const}, id_1^{const}$
$x^y + y^2, id_1^{id_2} + id_2^2,$
$x^y + y^{const}, id_1^{id_2} + id_2^{const}$
$x^y + y^{const}, id_1^{id_2} + id_2^{const}$
Match!

Math in DSpace (DML-CZ)

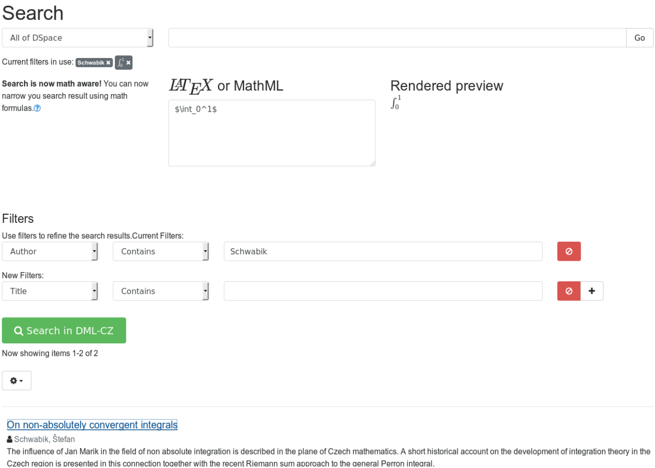
Assume we have already prepared MathML in time of ingest.

- extra metadata registry for math
    - *dmlcz.math*: MathML formulae storage
- configure SOLR to process *dmlcz.math*
    - we use *search* core
    - index *dmlcz.math* using **MIaS** analyzer
        - increases the size of index (approx. 100 times)
        - in DML-CZ: 264 MB → 28 GB (40 thousand items)
    - search *dmlcz.math* using **MIaS** analyzer again + **MIaS Payload Similarity** module
        - MIaS Payload Similarity takes care of results ranking

Math in DSpace (cont.)

- integrate 'user friendly' formulae search in DSpace UI
    - in our case XMLUI
- separate form for math search
    - MathML or LaTeX notation
    - on the fly rendered and displayed using **MathJax**
        - JavaScript library
    - LaTeX converted to MathML query via LaTeXML
        - written in Perl

Math formula search example in DML-CZ



Search

Current filters in use: Schwabik × $\int_0^1$ ×

Search is now math aware! You can now narrow you search result using math formulas.

LaTeX or MathML

```
$\int_0^1$
```

Rendered preview

$\int_0^1$

Filters

Use filters to refine the search results.Current Filters:

Author · Contains · Schwabik

New Filters:

Title · Contains

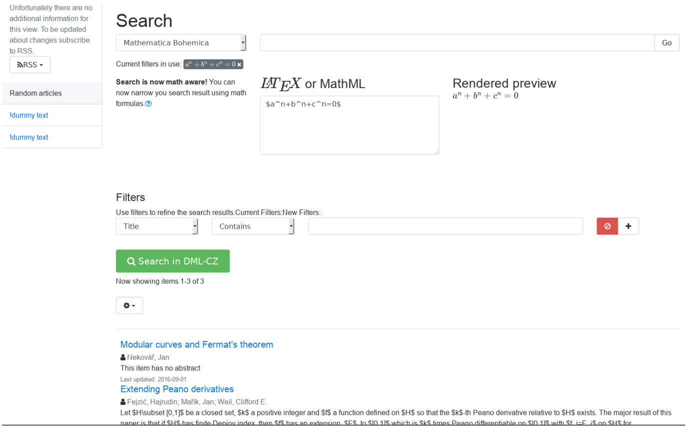
Now showing items 1-2 of 2

On non-absolutely convergent integrals

Schwabik, Štefan

The influence of Jan Marik in the field of non absolute integration is described in the plane of Czech mathematics. A short historical account on the development of integration theory in the Czech region is presented in this connection together with the recent Riemann sum approach to the general Perron integral.

Math formula search example in DML-CZ



Unfortunately there are no additional information for this view. To be updated about changes subscribe to RSS.

Random articles

!dummy text

!dummy text

Search

Mathematica Bohemica

Current filters in use: $a^n + b^n + c^n = 0$

Search is now math aware! You can now narrow you search result using math formulas.

LaTeX or MathML

```
$a^n+b^n+c^n=0$
```

Rendered preview

$a^n + b^n + c^n = 0$

Filters

Use filters to refine the search results.Current Filters:New Filters:

Title

Contains

Now showing items 1-3 of 3

Modular curves and Fermat's theorem

Nekovář, Jan

This item has no abstract

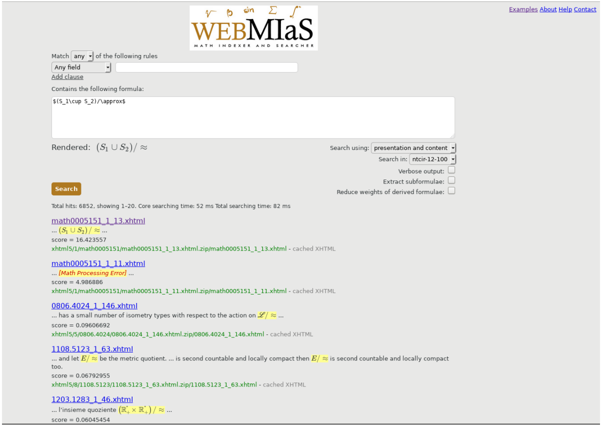
Last updated: 2016-09-01

Extending Peano derivatives

Fejzić, Hajrudin; Mařík, Jan; Weil, Clifford E.

Let $H\subset [0,1]$ be a closed set, k a positive integer and f a function defined on H so that the k-th Peano derivative relative to H exists. The major result of this paper is that if H has finite Denjoy index, then f has an extension, F, to $[0,1]$ which is k times Peano differentiable on $[0,1]$ with $f_i=F_i$ on H for

Math formula search example in WebMIaS



Match any of the following rules

Any field

Contains the following formula:

```
$(S_1\cup S_2)/\approx$
```

Rendered: $(S_1 \cup S_2)/ \approx$

Search using: presentation and content

Search in: ntcir-12-100

Verbose output: □

Extract subformulae: □

Reduce weights of derived formulae: □

Total hits: 6852, showing 1–20. Core searching time: 52 ms Total searching time: 82 ms

math0005151_1_13.xhtml
… $(S_1 \cup S_2)/ \approx$ …
score = 16.423557
xhtml5/1/math0005151/math0005151_1_13.xhtml.zip/math0005151_1_13.xhtml - cached XHTML

math0005151_1_11.xhtml
… [Math Processing Error] …
score = 4.986886
xhtml5/1/math0005151/math0005151_1_11.xhtml.zip/math0005151_1_11.xhtml - cached XHTML

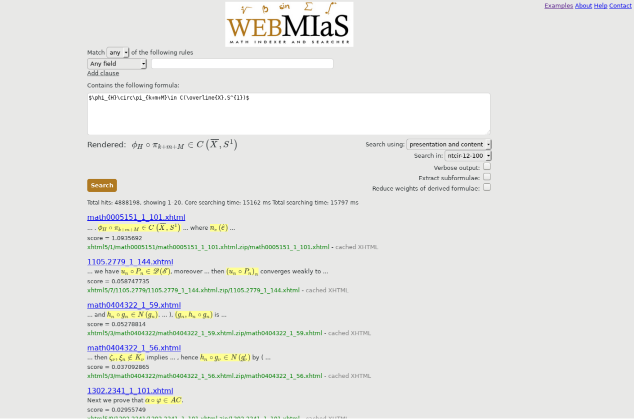
0806.4024_1_146.xhtml
… has a small number of isometry types with respect to the action on $\mathscr{L}/ \approx$ …
score = 0.09606692
xhtml5/5/0806.4024/0806.4024_1_146.xhtml.zip/0806.4024_1_146.xhtml - cached XHTML

1108.5123_1_63.xhtml
… and let $E/ \approx$ be the metric quotient. … is second countable and locally compact then $E/ \approx$ is second countable and locally compact too.
score = 0.06792955
xhtml5/8/1108.5123/1108.5123_1_63.xhtml.zip/1108.5123_1_63.xhtml - cached XHTML

1203.1283_1_46.xhtml
… l'insieme quoziente $(\mathbb{R}^*_+ \times \mathbb{R}^*_+)/ \approx$ …
score = 0.06045454

Data (MathML formulea) taken from ArXiv.org

Math formula search example in WebMIaS



Match any of the following rules

Any field

Contains the following formula:

```
$\phi_{H}\circ\pi_{k+m+M}\in C(\overline{X},S^{1})$
```

Rendered: $\phi_H \circ \pi_{k+m+M} \in C\left(\overline{X}, S^1\right)$

Search using: presentation and content

Search in: ntcir-12-100

Verbose output: □

Extract subformulae: □

Reduce weights of derived formulae: □

Total hits: 4888198, showing 1–20. Core searching time: 15162 ms Total searching time: 15797 ms

math0005151_1_101.xhtml
... , $\phi_H \circ \pi_{k+m+M} \in C\left(\overline{X}, S^1\right)$ ... where $n_e\left(\hat{e}\right)$ ...
score = 1.0935692
xhtml5/1/math0005151/math0005151_1_101.xhtml.zip/math0005151_1_101.xhtml - cached XHTML

1105.2779_1_144.xhtml
... we have $u_n \circ P_n \in \mathscr{D}(\mathscr{E})$, moreover ... then $(u_n \circ P_n)_n$ converges weakly to ...
score = 0.058747735
xhtml5/7/1105.2779/1105.2779_1_144.xhtml.zip/1105.2779_1_144.xhtml - cached XHTML

math0404322_1_59.xhtml
... and $h_n \circ g_n \in N\left(g_n\right)$. ...), $\left(g_n, h_n \circ g_n\right)$ is ...
score = 0.05278814
xhtml5/3/math0404322/math0404322_1_59.xhtml.zip/math0404322_1_59.xhtml - cached XHTML

math0404322_1_56.xhtml
... then $\zeta_\nu, \xi_n \notin K_\nu$ implies ... , hence $h_n \circ g_\nu \in N\left(g'_\nu\right)$ by (...
score = 0.037092865
xhtml5/3/math0404322/math0404322_1_56.xhtml.zip/math0404322_1_56.xhtml - cached XHTML

1302.2341_1_101.xhtml
Next we prove that $\alpha \circ \varphi \in AC$.
score = 0.02955749
xhtml5/9/1302.2341/1302.2341_1_101.xhtml.zip/1302.2341_1_101.xhtml - cached XHTML

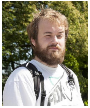
Data (MathML formulea) taken from ArXiv.org

Special thanks to (in alphabetical order):

Martin Líška (Maths Information Retrieval team)

Michal Růžička (Maths Information Retrieval team)

Petr Sojka (Maths Information Retrieval team)



Dominik Szalai (DSpace integration)

References

MIR team homepage:
`https://mir.fi.muni.cz/`

DML-CZ DSpace at GitHub (branch *dspace5-dmlcz*):
`https://github.com/empt-ak/DSpace`

DSpace and MIaS integration tech report:
`https://empt-ak.gitbooks.io/dmlcz/content/dml.html`

Questions?