Kanonizace MathML pro vyhledávání matematiky

Michal Růžička

<mruzicka@mail.muni.cz>

NLP seminář

4. května 2011

Motivace

- Index matematických výrazů v prohledávaných dokumentech je budován z MathML tokenizerem implementovaným v jazyce Java pomocí Apache Lucene/Solr.
- Konkrétní MathML matematický výraz je možno zapsat více způsoby.
  - $\Rightarrow$ Pro dosažení optimálních výsledků při vyhledávání je nutné sjednotit zápis MathML.

Motivace

- Index matematických výrazů v prohledávaných dokumentech je budován z MathML tokenizerem implementovaným v jazyce Java pomocí Apache Lucene/Solr.
- Konkrétní MathML matematický výraz je možno zapsat více způsoby.
  - $\Rightarrow$ Pro dosažení optimálních výsledků při vyhledávání je nutné sjednotit zápis MathML.

Motivace

- Index matematických výrazů v prohledávaných dokumentech je budován z MathML tokenizerem implementovaným v jazyce Java pomocí Apache Lucene/Solr.
- Konkrétní MathML matematický výraz je možno zapsat více způsoby.
  - $\Rightarrow$ Pro dosažení optimálních výsledků při vyhledávání je nutné sjednotit zápis MathML.

Kanonické MathML

- Matematické výrazy ve vstupních dokumentech i dotazy uživatelů jsou převedeny do „Canonical MathML“.
- Vycházíme ze sady XSLT, která je součástí nástroje UMCL – Universal Maths Conversion Library (<http://inova.ufr-info-p6.jussieu.fr/maths/umcl>).
  - Hlavním účelem nástroje UMCL je převod v MathML zapsaných matematických výrazů do národních variant Braillova písma.

Kanonické MathML

- Matematické výrazy ve vstupních dokumentech i dotazy uživatelů jsou převedeny do „Canonical MathML“.
- Vycházíme ze sady XSLT, která je součástí nástroje UMCL – Universal Maths Conversion Library (<http://inova.ufr-info-p6.jussieu.fr/maths/umcl>).
  - Hlavním účelem nástroje UMCL je převod v MathML zapsaných matematických výrazů do národních variant Braillova písma.

Kanonické MathML

- Matematické výrazy ve vstupních dokumentech i dotazy uživatelů jsou převedeny do „Canonical MathML“.
- Vycházíme ze sady XSLT, která je součástí nástroje UMCL – Universal Maths Conversion Library (<http://inova.ufr-info-p6.jussieu.fr/maths/umcl>).
  - Hlavním účelem nástroje UMCL je převod v MathML zapsaných matematických výrazů do národních variant Braillova písma.

Kanonické MathML (pokrač.)

- Pro použití ve WebMIaS provedeny drobné modifikace stylesheetů.
  - Odstraněno generování atributů `id="formula:xx"` pro každý uzel výstupního MathML.
  - Integrace transformace do budování indexu a zpracování dotazů.

Kanonické MathML (pokrač.)

- Pro použití ve WebMIaS provedeny drobné modifikace stylesheetů.
  - Odstraněno generování atributů `id="formula:`xx`"` pro každý uzel výstupního MathML.
  - Integrace transformace do budování indexu a zpracování dotazů.

Kanonické MathML (pokrač.)

- Pro použití ve WebMIaS provedeny drobné modifikace stylesheetů.
  - Odstraněno generování atributů `id="formula:xx"` pro každý uzel výstupního MathML.
  - Integrace transformace do budování indexu a zpracování dotazů.

Potřeba kanonizace MathML

- Kanonizace MathML zvyšuje přesnost vyhledávání.
- Pomáhá najít výsledky i pro dotazy, které by bez použití kanonizace neuspěly.

Potřeba kanonizace MathML

- Kanonizace MathML zvyšuje přesnost vyhledávání.
- Pomáhá najít výsledky i pro dotazy, které by bez použití kanonizace neuspěly.

Potřeba kanonizace MathML (pokrač.)

- Dotaz na formuli $x^2 + y^2$ zadanou MathML kódem:

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
    <msup>
        <mi>x</mi>
        <mn>2</mn>
    </msup>
    <mo>+</mo>
    <msup>
        <mi>y</mi>
        <mn>2</mn>
    </msup>
</math>
```

- Bez kanonizace není systém schopen najít žádné podobné formule, protože chybí element `<mrow>`.

Potřeba kanonizace MathML (pokrač.)

- Dotaz na formuli $x^2 + y^2$ zadanou MathML kódem:

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
    <msup>
        <mi>x</mi>
        <mn>2</mn>
    </msup>
    <mo>+</mo>
    <msup>
        <mi>y</mi>
        <mn>2</mn>
    </msup>
</math>
```

- Bez kanonizace není systém schopen najít žádné podobné formule, protože chybí element `<mrow>`.

Potřeba kanonizace MathML (pokrač.)

- Při použití kanonizace je dotaz převeden do následující podoby:

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <msup>
        <mi>x</mi><mn>2</mn></msup>
    <mo>+</mo>
    <msup>
        <mi>y</mi><mn>2</mn></msup>
  </mrow>
</math>
```

- Výsledkem je 36 817 zásahů v MREC 2011.4.

Potřeba kanonizace MathML (pokrač.)

- Při použití kanonizace je dotaz převeden do následující podoby:

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <msup>
        <mi>x</mi><mn>2</mn></msup>
    <mo>+</mo>
    <msup>
        <mi>y</mi><mn>2</mn></msup>
  </mrow>
</math>
```

- Výsledkem je 36 817 zásahů v MREC 2011.4.

Zpracování dotazů v LATEXu

- MathML syntax je složitá, hodí se především pro Copy & Paste výstupu automatizovaného nástroje.
- WebMIaS má integrovaný převodník z LATEXu do MathML.
  - Používá se Tralics.

Zpracování dotazů v LATEXu

- MathML syntax je složitá, hodí se především pro Copy & Paste výstupu automatizovaného nástroje.
- WebMIaS má integrovaný převodník z LATEXu do MathML.
  - Používá se Tralics.

Demo

Demo:

<http://aura.fi.muni.cz:8085/webmias/>

D. Archambault, V. Moço.

Canonical MathML to Simplify Conversion of MathML to Braille Mathematical Notations.

In K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, editors, *Computers Helping People with Special Needs*, volume 4061 of *Lecture Notes in Computer Science*, pages 1191–1198. Springer Berlin / Heidelberg, 2006.

<http://dx.doi.org/10.1007/11788713_172>.

J. Grimm.

Producing MathML with Tralics.

In Sojka [5], pages 105–117.

<http://dml.cz/dmlcz/702579>.

O. Kováčik, J. Rákosník.

On spaces $L^{p(x)}$ and $W^{k,p(x)}$.

Czechoslovak Mathematical Journal, 41:592–618, 1991.

<http://dml.cz/handle/10338.dmlcz/102493>.

B. R. Rowe, D. W. Wood, A. N. Link, D. A. Simoni.

Economic impact Assessment of NIST's Text REtrieval Conference (TREC) Program.

Technical Report RTI Project Number 0211875, July 2010.

<http://trec.nist.gov/pubs/2010.economic.impact.pdf>.

P. Sojka, editor.

Towards a Digital Mathematics Library, Paris, France, July 2010. Masaryk University.

<http://www.fi.muni.cz/ sojka/dml-2010-program.html>.

P. Sojka, M. Líška.

The Art of Mathematics Search, Apr. 2011.

Submitted to DocEng 2011 as full paper.

H. Stamerjohanns, D. Ginev, C. David, D. Misev, V. Zamdzhiev, M. Kohlhase.

MathML-aware Article Conversion from LaTeX.

In P. Sojka, editor, *Proceedings of DML 2009*, pages 109–120, Grand Bend, Ontario, CA, July 2009. Masaryk University.

<http://dml.cz/dmlcz/702561>.

H. Stamerjohanns, M. Kohlhase, D. Ginev, C. David, B. Miller.

Transforming Large Collections of Scientific Publications to XML.

Mathematics in Computer Science, 3:299–307, 2010.

<http://dx.doi.org/10.1007/s11786-010-0024-7>.

W. Sylwestrzak, J. Borbinha, T. Bouche, A. Nowiński, P. Sojka.

EuDML—Towards the European Digital Mathematics Library.

In Sojka [5], pages 11–24.

<http://dml.cz/dmlcz/702569>.