Mathematical Indexing and Querying

Martin Líška et al.

Masaryk University, Faculty of Informatics
Brno, Czech Republic
<255768@mail.muni.cz>

4th May, 2011
Motivation

• European Digital Mathematics Library – mathematical searching *should* be provided

• Conventional searching approaches are not applicable

• WP 5 Metadata repository and search engine implementation

• Deliverables D5.2 – The EuDML Search and Browsing Service (Demo due in M12, 18 PMs) and D5.3 (final, in M30, 26 PMs)

<table>
<thead>
<tr>
<th>Task</th>
<th>IST</th>
<th>FIZ</th>
<th>MU</th>
<th>ICM</th>
<th>Total</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>9</td>
<td>11</td>
<td></td>
<td>M04</td>
<td>M12</td>
</tr>
<tr>
<td>T5.3 Searching and browsing – simple version</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5.4 Searching and browsing – advanced version</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>13</td>
<td></td>
<td>M09</td>
<td>M24</td>
</tr>
<tr>
<td>Total PMs per WP5</td>
<td>24</td>
<td>10</td>
<td>10</td>
<td>24</td>
<td>68</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Math Indexer and Searcher – Features

- Based on full-text core Apache Lucene
- Presentation MathML
- Allows similarity (not only exact match) between query and matched term
  - Commutativity
  - Unification of variables and number constants
  - Subformulae matching
- Level of similarity calculation for expressions
- Mixed mathematical-textual queries
- Match snippet generation
Math Indexer and Searcher – Design

Motivation
Math Indexer and Searcher
Evaluation
Conclusion

Mathematical Indexing and Querying
Masaryk University, Faculty of Informatics, Brno, Czech Republic
Math Indexer and Searcher – Design II

Mathematical Indexing and Querying

Masaryk University, Faculty of Informatics, Brno, Czech Republic
Math Indexer and Searcher – Weighting

- We used the weighting utility

- Indexing
  - initial weight $= \frac{1}{\text{number_of_nodes}}$
  - level coefficient $l = 0.7$
  - variables coefficient $v = 0.8$
  - constants coefficient $c = 0.5$

- Searching
  - $\text{result} \times \text{number_of_query_nodes}$
Mathematical Indexing and Querying

Masaryk University, Faculty of Informatics, Brno, Czech Republic

**Formula Processing Example**

**Motivation**

**Math Indexer and Searcher**

**Evaluation**

**Conclusion**

```
input:
arranged:
tokenization:
variables unification:
constants unification:

(a + b^2 + c, 0.125)

("mi" < "mn" ⇒ 2 <-> c)

(a + b^{c+2}, 0.125)

(a, 0.0875)

(b, 0.06125)

(c + 2, 0.06125)

(b c + 2, 0.0875)

(c, 0.042875)

(c + 2, 0.042875)

(+, 0.042875)

(+, 0.0875)

(b^{c+2}, 0.0875)

(id_1 + id_2^{id_1 + 2}, 0.1)

(id_{id_1 + 2}, 0.07)

(id_{id_1 + 2}, 0.033)

(id_{id_1 + 2}, 0.035)

(id_{id_1 + const}, 0.030625)

(id_{id_1 + const}, 0.01715)

(a + b^{c + \text{const}}, 0.0625)

(a + b^{c + \text{const}}, 0.05)

(b^{c + \text{const}}, 0.04375)

(id_{id_1 + id_2^{id_1 + const}}, 0.05)

(id_{id_1 + id_2^{id_1 + const}}, 0.05)

(id_{id_1 + id_2^{id_1 + const}}, 0.05)

(id_{id_1 + id_2^{id_1 + const}}, 0.05)

(id_{id_1 + const}, 0.030625)

(id_{id_1 + const}, 0.01715)

(a + b^{c + \text{const}}, 0.0625)

(a + b^{c + \text{const}}, 0.05)

(b^{c + \text{const}}, 0.04375)
```

Mathematical Indexing and Querying

Masaryk University, Faculty of Informatics, Brno, Czech Republic
Implementation

- Java
- Lucene 3.1.0
- Mathematical part implements Lucene’s interface Tokenizer – able to integrate to any Lucene based system
  - MIaS4Solr plugin was created for the use in SOLR in EuDML
Evaluation

• MREC 2011.3.324
  • 324 060 documents
  • Uncompressed size 53 GB, compressed 6.3 GB
  • 112 million input formulae, over 2 billion expressions indexed
  • Index size 45 GB
  • Download here

• MREC 2011.4.439
  • 439 423 documents
  • Uncompressed size 124 GB, compressed 15 GB
  • 158 million input formulae, 2.9 billion expressions indexed
  • Index size 63 GB
  • Download here
WebMIaS

• Demo web interface: WebMIaS
  • MathML/TEX input
  • Canonization of the query – UMCL
  • Matched document snippet generation
Conclusion

MIaS is

- *text+math IR compatible* (fits mathematician’s needs)
- *scalable* (index with almost 3 billions formulae tested)
- *Lucene/SOLR compatible* system
Conclusion

• Project pages – MIaS/WebMIaS

• Future work
  • Reindex new corpus with canonized mathematics
  • Optimization
  • Mathematical equivalence computation via symbolic algebra system?
  • Suggestions welcomed
Questions?