Leveraging Mathematical Subject Information to Enhance Bibliometric Data

Maria Koutrakí, Olaf Teschke, Harald Sack, Fabian Müller, and Adam Bannister

28 May 2017, Portoroz Slovenia
Scientometrics@eswc2017
Agenda

- Introduction
  - Challenges in Mathematics
  - MSC classification system

- Approach
  - zbMATH data
  - Analysis

- Findings

- Conclusions
Why mathematics is challenging for bibliometrics?

1. Few publications and references → metrics are sensitive to distortions
2. The unusual longevity of mathematical research
3. The diverse nature of mathematics
   → Quantitative methods like bibliometric analysis are prone to misrepresentations!

Need to consider subject information!!
Mathematical Subject Classification (MSC)

- Mathematical Subject Classification (MSC)
  - Classification scheme maintained by Mathematical Reviews, zbMATH
  - Introduced in 1970
  - Revised every 10 years
  - Current version: MSC2010
  - Linked to DBpedia
  - Three-level classification tree
    - 63 first-level nodes
    - >400 second-level nodes
    - >5000 leaf nodes
zbMATH Data

- Zentralblatt MATH (zbMATH) [1]: service for abstracting and reviewing pure and applied mathematical documents.
- Contains >3.5 M bibliographic entries.
- Coverage from 18th century up to now!
- Available online

[1]: https://zbmath.org/

zbMATH

Free access is limited to 3 results, and filter functions are disabled. For full access subscription is required.

Found 145631 documents (Results 1–100)

Zimmermann, Wolf; Picht, Roswitha
(Konzepte höherer Programmiersprachen. (to appear).) (German) [Zbl 05834957]
MSC: 68N15 68N01 68-01
Statistical Analysis of zbMATH Data (I)

Distribution of zbMATH papers and authors in the different MCS categories.
Citation frequency of top level MSC categories.

Highly diverse citation frequency for the top level MSC categories.

→ Necessity to consider subject information for bibliometric studies in mathematics.
Basic Hierarchy of MSC2010 (SKOS Core)

```ttl
msc2010:53A40 a skos:Concept;
    skos:inScheme msc:2010: ;
    skos:broader msc:53Axx ;
    skos:preferredLabel "Other special differential geometries@en" ;
```
MSC2010 links to DBpedia [Lange et al. CICM’12]

970 MSC2010 subject linked to 2,960 DBpedia entities.
Adjustments on links to DBpedia

• Replace **redirects** with the URI of the original DBpedia entity.

• Replace the URIs of DBpedia entities *changed* from previous versions.

• Extension of links to DBpedia.
MSC2010 links to DBpedia -- Extended
MSC2010 links to DBpedia -- Extended

New links
+ 1,525 skos:narrower
+ 234 skos:broader
+ 283 skos:seeAlso

Concept-Scheme

Concept

Concept

Concept

DBpedia-Entity

DBpedia-Entity
Similarity of MSC2010 Categories

• **Semantic similarity:** wr.t. DBpedia entities corresponding to MSC2010 categories.
  • Use similarity method proposed in [Piao et al.]
  • Properties of the resources
  • Other similarity measures: equal self-similarity, symmetry, minimality

• **Statistical similarity:** w.r.t MSC2010 categories assigned to zbMATH paper collection.
  • Symmetric
    \[
    J_{\text{accard}}(\text{cat}_a, \text{cat}_b) = \frac{\text{cat}_a \cap \text{cat}_b}{\text{cat}_a \cup \text{cat}_b}
    \]
  • Asymmetric
    \[
    J_{\text{accard}}(\text{cat}_a \rightarrow \text{cat}_b) = \frac{\text{cat}_a \cap \text{cat}_b}{\text{cat}_a}
    \]
Experimental Setup - Results

• **Statistical similarity**: zbMATH collection of papers a.k.a. 100K papers.

• **Semantic similarity**: Top 10K category pairs from statistical similarity.
  - Symmetric: 937 mapped to DBpedia via owl:sameAs.
  - Asymmetric: 672 mapped to DBpedia via owl:sameAs.

• Correlations coefficient of semantic and statistical similarities

<table>
<thead>
<tr>
<th>Similarity measures</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic – Jaccard</td>
<td>0.34</td>
</tr>
<tr>
<td>Semantic – asymmetric Jaccard</td>
<td>-0.04</td>
</tr>
</tbody>
</table>
Issues in MSC2010 (1/2)

- **Structural Issues**

<table>
<thead>
<tr>
<th>Super-category</th>
<th>Sub-category</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>13Gxx</td>
<td>13G05</td>
<td>Integral domains</td>
</tr>
<tr>
<td>14Txx</td>
<td>14T05</td>
<td>Tropical geometry</td>
</tr>
<tr>
<td>22Cxx</td>
<td>22C05</td>
<td>Compact groups</td>
</tr>
<tr>
<td>45Qxx</td>
<td>45Q05</td>
<td>Inverse problems</td>
</tr>
<tr>
<td>62Qxx</td>
<td>62Q05</td>
<td>Statistical tables</td>
</tr>
<tr>
<td>65Axx</td>
<td>65A05</td>
<td>Tables</td>
</tr>
<tr>
<td>85-XX</td>
<td>85Axx</td>
<td>Astronomy and astrophysics</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Issues in MSC2010 (2/2)

• Imprecise `owl:sameAs` links
  • Links to `redirects` resources to DBpedia.

• **Erroneous** links between the MSC2010 and DBpedia.
  • E.g.:

```latex
\begin{tikzpicture}
  \node [shape=circle,draw=blue, style={fill=blue!20}] (A) {msc:65A05};
  \node [shape=circle,draw=blue, style={fill=blue!20}] (B) at (2,0) {dbr:tablets};
  \draw [->, blue] (A) -- (B) node [midway, above] {`owl:sameAs`};
  \node [rectangle, draw=blue, style={fill=blue!20}] (C) at (0,-2) {skos:preferedLabel};
  \node [rectangle, draw=blue, style={fill=blue!20}] (D) at (0,-3) {``tables@en``};
  \draw [->, blue] (C) -- (D);
\end{tikzpicture}
```
• Imprecise `owl:sameAs` links
  • Links to `redirects` resources to DBpedia.

• **Erroneous** links between the MSC2010 and DBpedia.
  • E.g.:

```
msc:65A05
  skos:preferedLabel
    “tables@en”
  owl:sameAs
    "tables@en"

owl:sameAs

dbr:tables

owl:sameAs

dbr:tables
```
Conclusions

- Correct, upgrade, and extend existing MSC2010 mappings to DBpedia.

- Computation of MSC categories similarity through statistical and semantic measures.

- Detection of inconsistencies in MSC2010.
Questions?

Thank you!

We are hiring!
Come and talk to us!
ISE group @AIFB KIT