Same Data, Different Results-- On a Comparative Topic Extraction Exercise

Theresa Velden
University of Michigan School of Information (UMSI)

SIGMET Workshop at ASIST 2015
November 7, 2015
In collaboration with:

Kevin Boyack (SciTech Strategies) · Nees van Eck (CWTS Leiden) · Wolfgang Glänzel & Bart Thijs (ECOOM) · Jochen Gläser (TU Berlin) · Frank Havemann & Michael Heinz (HU Berlin) · Rob Koopman & Shenghui Wang (OCLC Research), Andrea Scharnhorst (DANS-KNAW)
The performative nature of topic extraction

• To what extent do topic extraction approaches capture the ‘ground truth’ of thematic structure in a field or how does the choice of approach shape the results and introduce artifactual features?

• In Scientometrics topic extraction approaches are rarely directly compared on same data set; lack of understanding of nature & origin, and implications of differences
Background

• Evolved from annual meetings of advisory project funded by German Ministry for Education and Research on ‘Measuring Diversity in Science’ (Jochen Gläser, Frank Havemann & Michael Heinz)
• To measure epistemic diversity of a field, the field needs to be delineated and topics identified
  – Even slight changes in topic structure influence measure
• Compare solutions derived from same data set (‘Astro Data’)
• Special session at ISSI 2015, July in Istanbul
Premises & Objective

• More than one valid thematic structure can be constructed depending on the perspective applied to the knowledge.

• Topical structures are reconstructed for specific purposes, so if at all, there might be a best method for a given purpose.

• Instead of finding the one best solution, we aim at uncovering how results differ and how those differences relate to approaches
Data Set

• Source: Web of Science (Thomson Reuters)
• 8 years: 2003 - 2010
• 59 astrophysics and astronomy journals
• 111,161 articles, letters & proceedings papers
Topic Extraction Workflow

Raw data

Data Cleaning
- Matching with in-house database
- Citation cleaning

Data Model
- Bibliographic coupling
- Direct Citation
- Hybrid (bc & NLP)
- Semantic Matrix (incl. normalizations and thresholds)

Clustering Algorithm
- Louvian
- Infomap
- SLMA
- k-means
- memetic (incl. resolution and parameter settings)

Solution
- (sets of documents)
## Overview Approaches

<table>
<thead>
<tr>
<th></th>
<th>Direct Citation</th>
<th>Bibliogr. Coupling</th>
<th>Hybrid (bc &amp; terms/NLP)</th>
<th>Semantic matrix</th>
<th>Projection onto Global Direct Citation Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infomap</td>
<td>UMSI</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>SLMA</td>
<td>CWTS</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>STS</td>
</tr>
<tr>
<td>Memetic</td>
<td>HU</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Louvian</td>
<td>--</td>
<td>ECOOM</td>
<td>ECOOM</td>
<td>OCLC</td>
<td>--</td>
</tr>
<tr>
<td>K-means</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>OCLC</td>
<td>--</td>
</tr>
</tbody>
</table>

**HU:** Humboldt University; **CWTS:** Centre for Science and Technology Studies, Leiden; **ECOOM:** Expertisecentrum Onderzoek en Ontwikkelingsmonitoring; **UMSI:** University of Michigan School of Information, **OCLC:** Online Computer Library Center, Inc.; **STS:** SciTech Strategies
Results:
Cluster (topic) size distributions

![Graph showing cumulative fraction over cluster rank for different distributions.]

- SR(555)
- C(22)
- U(22)
- HD(71)*
- OK(31)
- OL(32)
- EB(13)
- EN(11)
Results

- Propor6on of unique paper pairs
- Concentration Index (Herfindahl := sum of squared topic sizes)

% of Unique Pairs

Concentration Index

y = 3.5936x - 0.1144
R² = 0.91194
Astroparticle Physics (6,8)

Gravitational Physics, Cosmology (2,13,14)

Astrophysics (Galaxies, Stars) (1,3,7,9-12,15,16)

Solar Physics (4)

Space Science (17,18,21,22)

Planetary Science (5,19,20)

Solution: UMS10 (direct citation & infomap)
Network vis: gephi, Force Atlas 2 Layout algorithm
Labeling: based on journal signature
Sources of variability

Same model & same algorithm (stochastic variation)

Same model & different algorithm

Different model & different algorithm

Overlap Between Clusters: Comparison with UMSI0 Cluster Solution (22 clusters)
## Comparison: Set based metrics

### Normalised Mutual Information

<table>
<thead>
<tr>
<th></th>
<th>sr</th>
<th>c</th>
<th>u</th>
<th>ok</th>
<th>ol</th>
<th>eh</th>
<th>eb</th>
</tr>
</thead>
<tbody>
<tr>
<td>sr</td>
<td>1.000</td>
<td>0.359</td>
<td>0.372</td>
<td>0.329</td>
<td>0.333</td>
<td><strong>0.243</strong></td>
<td><strong>0.306</strong></td>
</tr>
<tr>
<td>c</td>
<td>0.359</td>
<td>1.000</td>
<td>0.633</td>
<td>0.464</td>
<td>0.516</td>
<td>0.316</td>
<td>0.380</td>
</tr>
<tr>
<td>u</td>
<td>0.372</td>
<td><strong>0.633</strong></td>
<td>1.000</td>
<td>0.424</td>
<td>0.471</td>
<td><strong>0.295</strong></td>
<td>0.356</td>
</tr>
<tr>
<td>ok</td>
<td>0.329</td>
<td>0.464</td>
<td>0.424</td>
<td>1.000</td>
<td><strong>0.515</strong></td>
<td>0.334</td>
<td>0.363</td>
</tr>
<tr>
<td>ol</td>
<td>0.333</td>
<td><strong>0.516</strong></td>
<td>0.471</td>
<td>0.515</td>
<td>1.000</td>
<td>0.307</td>
<td>0.362</td>
</tr>
<tr>
<td>eh</td>
<td>0.243</td>
<td>0.316</td>
<td>0.295</td>
<td>0.334</td>
<td>0.307</td>
<td>1.000</td>
<td>0.330</td>
</tr>
<tr>
<td>eb</td>
<td>0.306</td>
<td>0.380</td>
<td>0.356</td>
<td>0.363</td>
<td>0.362</td>
<td>0.330</td>
<td>1.000</td>
</tr>
</tbody>
</table>

### Overlap Index

<table>
<thead>
<tr>
<th></th>
<th>sr</th>
<th>c</th>
<th>u</th>
<th>ok</th>
<th>ol</th>
<th>eh</th>
<th>eb</th>
</tr>
</thead>
<tbody>
<tr>
<td>sr</td>
<td>1.000</td>
<td>0.698</td>
<td>0.706</td>
<td>0.686</td>
<td>0.692</td>
<td>0.591</td>
<td>0.662</td>
</tr>
<tr>
<td>c</td>
<td>0.698</td>
<td>1.000</td>
<td><strong>0.835</strong></td>
<td>0.622</td>
<td><strong>0.708</strong></td>
<td><strong>0.546</strong></td>
<td>0.593</td>
</tr>
<tr>
<td>u</td>
<td>0.706</td>
<td><strong>0.835</strong></td>
<td>1.000</td>
<td>0.645</td>
<td><strong>0.725</strong></td>
<td><strong>0.526</strong></td>
<td>0.574</td>
</tr>
<tr>
<td>ok</td>
<td>0.686</td>
<td>0.622</td>
<td>0.645</td>
<td>1.000</td>
<td>0.619</td>
<td>0.609</td>
<td>0.576</td>
</tr>
<tr>
<td>ol</td>
<td>0.692</td>
<td>0.708</td>
<td>0.725</td>
<td>0.619</td>
<td>1.000</td>
<td>0.553</td>
<td>0.567</td>
</tr>
<tr>
<td>eh</td>
<td>0.591</td>
<td>0.546</td>
<td>0.526</td>
<td>0.609</td>
<td>0.553</td>
<td>1.000</td>
<td><strong>0.541</strong></td>
</tr>
<tr>
<td>eb</td>
<td>0.662</td>
<td>0.593</td>
<td>0.574</td>
<td>0.576</td>
<td>0.567</td>
<td>0.541</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Outside of the giant component: Blind spot of a citation based approach
Two very different solutions
Two very different solutions

UMSI0 (Direct citation & Infomap)

ECOOM-HY (bibliographic coupling/NLP term extraction & Louvain)

Network vis: gephi, Force Atlas 2 Layout algorithm
Labeling: based on Little Ariadne related words/subjects
Conclusions & Outlook

• Developing methods for meaningful comparison a major challenge
  – combination of quantitative metrics & visualizations
• Variations due to coverage, modeling & clustering
• Comparative analysis ongoing:
  – Case studies (instances of agreement and divergence)
  – Blind spots (areas left out by some approaches)
  – Mapping onto Unified Astronomy Thesaurus
• Special Issue in Scientometrics in preparation
• Call to join ‘Topic Extraction Challenge’

Theresa Velden, tvelden@umich.edu, University of Michigan School of Information
Comparison of UMSI0 and CWTS-C5

Visualization: Little Ariadne (OCLC)