Science and Technology in digital newspapers

Carlos G. Figuerola, Tamar Groves, Miguel Angel Quintanilla - ECyT Institute
University of Salamanca
II Seminar on Indicators of Scientific and Technological Culture - 25/11/2014
Digital Newspapers

- not as physical newspapers
- heterogeneous formats
- heterogeneous web site structures
- concerns with digital preservation
Digital Newspapers

- Three newspapers: El Mundo, El País, Público
- Time period: 2002-2011 (except Público, only since 2007)
- More than 900,000 news
Automatic Categorization
We are only interested on news about Science & Technology

- we can use an automatic supervised classifier
- SVM is a good choice
- we can try also SVM to classify news in the categories of our theoretic model

Training Process

- an initial sample built by hand
- an iterative process of classify - refining sample - retraining - reclassify
Results: the SCSC
50,753 news about S & T
More Results: Science vs. Technology

Science, Technology % on all news

- Sci
- Tech
- Sci + Tech

Intrinsec and extrinsec features

% intrinsec and extrinsec Sci
% intrinsec and extrinsec Tech

![Graph showing trends in intrinsec and extrinsec features over time from 2002 to 2011.](image)
Topics Discovering using SNA Techniques

- objects can establish relationships between them
- we can map objects and relationships towards a network or graph
  - objects are nodes
  - relationships are edges or links between nodes
Establishing relationships between news

- we can compute semantic similarity between documents
  - using borrowed techniques from the Information Retrieval field
  - applying the well known Vector Space Model
  - based on words and weights of each word inside each document
- news are nodes in a network
- there is an edge between two docs if they are similar
- the weight of this edge is the similarity's degree between both docs
Detecting Communities

- in a network, a community is a bunch of nodes
  - strongly linked between them
  - links weakly with nodes outside the bunch
- in our network of news, a community is a topic
- they are several algorithms to find communities in networks
- we use InfoMap: fast and efficient, accurate results
Analyzing Results

**Communities listing**

<table>
<thead>
<tr>
<th>community</th>
<th>topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Public Health</td>
</tr>
<tr>
<td>2</td>
<td>Biomedicine</td>
</tr>
<tr>
<td>3</td>
<td>Energy</td>
</tr>
<tr>
<td>4</td>
<td>Human Development</td>
</tr>
<tr>
<td>5</td>
<td>Natural Resources</td>
</tr>
<tr>
<td>6</td>
<td>Aerospace Research</td>
</tr>
<tr>
<td>7</td>
<td>Biodiversity</td>
</tr>
<tr>
<td>8</td>
<td>Astronomy &amp; Cosmology</td>
</tr>
<tr>
<td>9</td>
<td>Information Technology</td>
</tr>
<tr>
<td>10</td>
<td>Science Policy</td>
</tr>
<tr>
<td>11</td>
<td>Protected Species - Spain</td>
</tr>
<tr>
<td>12</td>
<td>Human Evolution</td>
</tr>
<tr>
<td>13</td>
<td>Contamination</td>
</tr>
</tbody>
</table>
## Analyzing Results

<table>
<thead>
<tr>
<th>Subcommunity</th>
<th>Topic</th>
<th>Subcommunity</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>influenza</td>
<td>1.11</td>
<td>infections, E. Coli,</td>
</tr>
<tr>
<td>1.2</td>
<td>AIDS</td>
<td>1.12</td>
<td>cholera</td>
</tr>
<tr>
<td>1.3</td>
<td>mortality</td>
<td>1.13</td>
<td>Legionella</td>
</tr>
<tr>
<td>1.4</td>
<td>drugs</td>
<td>1.14</td>
<td>polio</td>
</tr>
<tr>
<td>1.5</td>
<td>vaccines</td>
<td>1.15</td>
<td>mad cow disease</td>
</tr>
<tr>
<td>1.6</td>
<td>malaria</td>
<td>1.16</td>
<td>foot and mouth disease</td>
</tr>
<tr>
<td>1.7</td>
<td>SARS</td>
<td>1.17</td>
<td>dengue</td>
</tr>
<tr>
<td>1.8</td>
<td>tuberculosis</td>
<td>1.18</td>
<td>insect infections</td>
</tr>
<tr>
<td>1.9</td>
<td>hepatitis C</td>
<td>1.19</td>
<td>Chagas</td>
</tr>
<tr>
<td>1.10</td>
<td>antibiotics, bacteria</td>
<td>1.20</td>
<td>bio-bac</td>
</tr>
</tbody>
</table>
Conclusions

- more Sci than Tech
- in Sci news more intrinsically
- predominance of academic model of science communication
  - journalists tend to reproduce scientific information and they don't enter into questions of its social political or moral implications
- topics:
  - predominance of biomedicine
  - progressive growing of Information technologies
  - specific events produce punctual growth in news about ecology, pollution, ...
Conclusions: big data treatment

- We tried using automated information retrieval procedures to recuperate science news and several kinds of specialized software to classify and analyze it.
- Their usage was efficient in analyzing our vast corpus and reaching some preliminary conclusions.
- However we are left with the challenge of explaining the high number of unclassified articles related to our model.
- There is a need to analyze more carefully the sub clusters and their significance.
<Thank You!>

Important contact information goes here.

e-mail

www