Entity Resolution of Institutions in Bibliographic Databases

Jeffrey Fisher

Supervisors: Assoc Prof. Peter Christen, Dr. Qing Wang, Dr. Paul Wong

COMP 8740: AI Project, Semester 1, 2013
Australian National University
Overview

- Background and Definitions
- Data Exploration
- Data Cleaning
- Data Matching
- Conclusions
- Future Work
Background

- Databases such as SCOPUS Elsevier and Thomson Scientific contain bibliographic information including: publication information, conference records, authors and institutions.

- Applications:
  - Increasingly used to measure research performance and allocate government grant funding.
  - Can potentially be used to predict future research patterns and emerging growth areas. Can we spot the next hot topic in research?

- Applications such as these require clean data.
Entity Resolution

- Entity Resolution has two main aspects:
  - Determining which records or values correspond to the same real world entity and then combining them.
  - Determining where a single record or value corresponds to multiple real world entities and then separating it.

- Entity resolution can be conducted within a single database, or between two or more databases.
- For this project we used a single bibliographic database, and focused entirely on the first aspect. This is often referred to as deduplication.
Data Exploration

- The project was done on the SCOPUS (Elsevier) Database.
- SCOPUS stores the data in XML format.
- There is little normalisation of authors or institutions.
- SCOPUS contains approximately 80 million records. Of these approximately 1.6 million are for Australia.
## Example Data

Sample data from SCOPUS Bibliographic Database.

<table>
<thead>
<tr>
<th>afid</th>
<th>dptid</th>
<th>country</th>
<th>organization</th>
<th>city_group</th>
<th>address</th>
</tr>
</thead>
<tbody>
<tr>
<td>*96</td>
<td>aus</td>
<td>aus</td>
<td>Telethon Inst. for Child Hlth. Res., Faculty of Medicine and Dentistry, University of Western Australia</td>
<td></td>
<td>Perth, WA</td>
</tr>
<tr>
<td>*35</td>
<td>aus</td>
<td>aus</td>
<td>Equine Concepts</td>
<td></td>
<td>Mount Martha, VIC 3934</td>
</tr>
<tr>
<td>*73</td>
<td>aus</td>
<td>aus</td>
<td>Private Ophthalmology Practice</td>
<td></td>
<td>Spring Hill, QLD</td>
</tr>
<tr>
<td>*73</td>
<td>aus</td>
<td>aus</td>
<td>Private Ophthalmology Practice</td>
<td></td>
<td>Chermside, QLD</td>
</tr>
<tr>
<td>*06</td>
<td>aus</td>
<td>aus</td>
<td>Premion</td>
<td></td>
<td>Nambour 4560, QLD</td>
</tr>
<tr>
<td></td>
<td>aus</td>
<td>aus</td>
<td>Premion</td>
<td></td>
<td>10 Mapelton Road</td>
</tr>
<tr>
<td>*82</td>
<td>*62</td>
<td>aus</td>
<td>The Prince of Charles Hospital, Health Service District</td>
<td></td>
<td>Chermside, QLD</td>
</tr>
<tr>
<td>*82</td>
<td>*62</td>
<td>aus</td>
<td>The Prince of Charles Hospital, Health Service District</td>
<td></td>
<td>Chermside, QLD</td>
</tr>
<tr>
<td>*39</td>
<td>*97</td>
<td>aus</td>
<td>Department of Respiratory Medicine, Alfred Hospital</td>
<td></td>
<td>Sydney, NSW</td>
</tr>
<tr>
<td>*66</td>
<td>*20</td>
<td>aus</td>
<td>National Health and Medical Research Council (NHMRC), Centre of Clinical Research Excellence in Nutritional Physiology, Interventions and Outcomes</td>
<td></td>
<td>Adelaide, SA 5000</td>
</tr>
<tr>
<td>*27</td>
<td>aus</td>
<td>aus</td>
<td>Neptune Marine Services</td>
<td></td>
<td>Perth, WA</td>
</tr>
<tr>
<td>*81</td>
<td>aus</td>
<td>aus</td>
<td>Universidad Politécnica de Victoria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*81</td>
<td>aus</td>
<td>aus</td>
<td>Universidad Politécnica de Victoria</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Cleaning

- The problem:
  - 238,460 unique organization values
  - 28,306 unique afid values

- We limited the scope to institutions with 10 or more records leaving 2,910 unique afid values.

- We used a frequency based approach.
- Comma separated tokens.
- Ratio: ratio of frequencies between the most common token and second most common token.
Example Ratio Calculation

- Most frequent token: “Australian National University” frequency: 29,311.
- Second most frequent token: “Research School of Chemistry” frequency: 5,483.
- Ratio = 29,311 / 5,483 = 5.35

<table>
<thead>
<tr>
<th>Comma Separated Token</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian National University</td>
<td>29,311</td>
</tr>
<tr>
<td>Research School of Chemistry</td>
<td>5,483</td>
</tr>
<tr>
<td>Research School of Physical Sciences and Engineering</td>
<td>4,485</td>
</tr>
<tr>
<td>John Curtin School of Medical Research</td>
<td>4,447</td>
</tr>
</tbody>
</table>
Data Cleaning Results

- Analysis of the results based on 200 randomly sampled \textit{afid} values and the names we extracted.
- We scored them as correct, incorrect or partially correct. Partially correct usually meant the name was correct but had an acronym or abbreviation.

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Number of Records from Sample (Total 200)</th>
<th>Correct</th>
<th>Partially Correct</th>
<th>Incorrect</th>
<th>% Correct or Partially Correct</th>
<th>% Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio &gt;= 4.00</td>
<td>59</td>
<td>50</td>
<td>8</td>
<td>1</td>
<td>98.3%</td>
<td>1.7%</td>
</tr>
<tr>
<td>2.00 &lt;= Ratio &lt; 4.00</td>
<td>61</td>
<td>42</td>
<td>11</td>
<td>8</td>
<td>86.9%</td>
<td>13.1%</td>
</tr>
<tr>
<td>1.50 &lt;= Ratio &lt; 2.00</td>
<td>25</td>
<td>14</td>
<td>8</td>
<td>3</td>
<td>88.0%</td>
<td>12.0%</td>
</tr>
<tr>
<td>1.20 &lt;= Ratio &lt; 1.50</td>
<td>20</td>
<td>12</td>
<td>5</td>
<td>3</td>
<td>85.0%</td>
<td>15.0%</td>
</tr>
<tr>
<td>1.10 &lt;= Ratio &lt; 1.20</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>80.0%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Ratio &lt; 1.10</td>
<td>25</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>60.0%</td>
<td>40.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>200</strong></td>
<td><strong>131</strong></td>
<td><strong>42</strong></td>
<td><strong>27</strong></td>
<td><strong>86.50%</strong></td>
<td><strong>13.50%</strong></td>
</tr>
</tbody>
</table>
Data Matching

- The data matching was done in two parts:
  - matching between different *afid* values
  - matching records with no *afid* value against institutions.

- We used different string comparison techniques such as Q-gram, Longest Common Substring and Jaro comparison.

- There was no truth data, so precision was used for evaluation.
Methodology

- When matching between different *afid* values, we used a clustering approach and merged clusters when a match occurred.

- When matching records with no *afid*, we had 18,808 records with an *organization* value.

- For both parts we matched with the 2,910 names extracted during the data cleaning phase.
Results (Part 1)

- Started with 2,910 clusters:
  - Exact match:
    - 566 clusters merged to 230 new clusters with 85.9% pair-wise precision.
  - Q-gram match, similarity threshold = 0.9, Q = 3:
    - 75 clusters merged to 35 new clusters with 87.0% pair-wise precision.
- No easy next step. However a rough analysis suggested that we had found most of the true matches.
Results (Part 2)

- Exact matching: 5,822 records matched with a precision of 96%.

- Q-gram, similarity = 0.9, Q = 2: 1,815 records matched with a precision of 96%.

- The next techniques were not so successful with the best ones giving 60% precision.

- We found three main reasons for the low precision: similar institution names, incorrect names from data cleaning phase and missing names.
Conclusions

- Extracting institution names was quite successful with over 85% of institutions having a correct or partially correct name.
- Data matching between *afid* values was reasonably good. Precision was over 85% with reasonable coverage.
- Matching records with no *afid* was less successful. 40% coverage with over 95% precision but results dropped away quickly after this.
- We explored a few reasons why the matching against records with no *afid* was less successful.
Future Work

- Other countries.
- Abbreviations and acronyms.
- TF-IDF approach.
- External data sources.
- Expand the entity resolution to look at authors and papers as well.
Questions?

Jeffrey Fisher: u4814620@anu.edu.au

Thank You