Ranking Universities Using Linked Open Data
Rouzbeh Meymandpour and Joseph G. Davis

Knowledge Discovery and Management Research Group
School of Information Technologies
Introduction

University- and Research-Related Content on Linked Data

Ranking Methodology

Evaluation and Experiments

Discussions

Conclusion and Future Work
Introduction

Linked Data

Semantic Web technologies have enabled the

Web of Data
a.k.a.
Linked Data


University Ranking Problem
Introduction Cont.

Linked Open Data
University-Related Content on Linked Open Data

1st Level

- dbo:affiliation (in)
- dbo:numberOfPostgraduateStudents (out)
- dbo:almaMater (in)
- dbo:numberOfStudents (out)
- dbo:campus (out)
- dbo:numberOfUndergraduateStudents (out)
- dbo:chancellor (out)
- dbo:occupation (in)
- dbo:city (out)
- dbo:president (out)
- dbo:college (in)
- dbo:staff (out)
- dbo:education (in)
- dbo:team (in)
- dbo:employer (in)
- dbo:training (in)
- dbo:facultySize (out)
- dbo:viceChancellor (out)

2nd Level

- dbo:author (out)
- dbo:field (out)
- dbo:award (out)
- dbo:influenced (in/out)
- dbo:designer (out)
- dbo:knownFor (out)
- dbo:doctoralAdvisor (in/out)
- dbo:developer (out)
- dbo:notableStudent (in/out)
- dbo:doctoralStudent (in/out)
- dbo:notableWork (out)
- dbo:foundedBy (in)
- dbo:developer (out)
Informativeness Measurement

Information Content (IC)

The amount of binary symbols (bits) required in order to recreate the transmitted process

\[ IC(a) = -\log(\pi(a)) \]

- \( \pi(a) \): the probability of presence of concept \( a \) in its corpus
- Also known as Shannon’s Theory of Communication (1948)
Formal Definition of Linked Data

› Each resource is a set of its features
  - \( A = \{(l_1, c, out), (l_2, d, in), (l_3, e, out), (l_4, f, out)\} \)

› A resource is described using its relations with neighbors
  - Incoming and outgoing edges
  - Semantics (link types)
  - The Direction of Links
Partitioned Information Content (PIC)*

\[ IC(A) = - \log(\pi(A)) = - \log\left(\pi(a_1) \pi(a_2) \cdots \pi(a_{|A|})\right) \]

\[ PIC(A) = \sum_{\forall a_i \in A} IC(a_i) \]

\[ \pi(a_i) = \frac{\varphi(a_i)}{N} \]

\[ \varphi(a_i) \] is the frequency of the feature \( a_i \)

\[ N \] is the frequency of the most common feature

Characteristics of PIC

› A simple example:
  - University of Sydney: Located in Sydney, vs.
  - University of Sydney: Member of G8
Developing the Ranking Metric

› Adjusting the influence of each relation:

\[ WPI\mathcal{C}(F_r) = \sum_{\forall f_i \in F_r} w_i \ IC(f_i) \]

› Extracting semantics in deeper layers:

\[ WPI\mathcal{C}(F_r)_k = WPI\mathcal{C}(F_r) + \sum_{\forall f_i \in F_r} w_i \ WPI\mathcal{C}(F_{f_i})_{k-1} \]

\[ k > 1 \]
Ranking Methodology Cont.

PIC₁

University

PIC₂

...
Evaluation

Evaluation Context

- Dataset: DBpedia 3.8 (Aug 2012)
- Semi-automatic Control to eliminate redundancy and noise
  - ‘dbo:almaMater’ relations have to connect universities to a ‘dbo:Person’

<table>
<thead>
<tr>
<th>University (First Depth)</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:almaMater</td>
<td></td>
<td>dbo:president</td>
<td>1</td>
</tr>
<tr>
<td>dbo:education</td>
<td></td>
<td>dbo:chancellor</td>
<td>1</td>
</tr>
<tr>
<td>dbo:team</td>
<td></td>
<td>dbo:dean</td>
<td>1</td>
</tr>
<tr>
<td>dbo:team</td>
<td></td>
<td>dbo:viceChancellor</td>
<td>1</td>
</tr>
<tr>
<td>dbo:training</td>
<td></td>
<td>dbo:head</td>
<td>1</td>
</tr>
<tr>
<td>dbo:occupation</td>
<td></td>
<td>dbo:publisher</td>
<td>1</td>
</tr>
<tr>
<td>dbo:employer</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person (Second Depth)</th>
<th>4</th>
<th>2</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:award</td>
<td></td>
<td>dbo:keyPerson</td>
<td>2</td>
</tr>
<tr>
<td>dbo:knownFor</td>
<td>2</td>
<td>dbo:foundedBy</td>
<td>2</td>
</tr>
<tr>
<td>dbo:doctoralAdvisor</td>
<td>1</td>
<td>dbo:doctoralStudent</td>
<td>1</td>
</tr>
<tr>
<td>dbo:influenced</td>
<td>2</td>
<td>dbo:notableWork</td>
<td>2</td>
</tr>
<tr>
<td>dbo:notableStudent</td>
<td>2</td>
<td>dbo:designer</td>
<td>2</td>
</tr>
<tr>
<td>dbo:author</td>
<td>2</td>
<td>dbo:developer</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication (Second Depth)</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:academicDiscipline</td>
<td></td>
<td>dbo:author</td>
</tr>
<tr>
<td>dbo:editor</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Evaluating Metrics

- Simple PIC-based Ranking Metric (PIC(Basic))
  - Only considers immediate neighbours
  - Without any weightings
  - All kinds of links without any restriction or control

- 2-Level PIC-based Ranking Metric (PIC)

Evaluating against:
- QS World University Rankings (QS)
- THE World University Rankings (THE)
- SJTU Academic Ranking of World Universities (SJTU)
Evaluation Cont.

Evaluation Metrics

1. Correlation of Scores
   - Matched the universities in each list with their corresponding DBpedia URI
     - Pearson Correlation Coefficient
     - Spearman Rank Correlation Coefficient

2. Similarity of top 100 lists
   - A list of 500 universities were chosen that includes all universities in all rankings
     (493 from QS + 7 missing universities)
     - Overlap Similarity
     - Average Overlap Similarity
       - Top-weighted (top of the rankings are more important)
## The Rankings*

<table>
<thead>
<tr>
<th>Rank</th>
<th>University</th>
<th>SJTU</th>
<th>QS</th>
<th>THE</th>
<th>PIC Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Harvard University</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>125,979.3</td>
</tr>
<tr>
<td>2</td>
<td>University of Cambridge</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>115,418.5</td>
</tr>
<tr>
<td>3</td>
<td>Princeton University</td>
<td>7</td>
<td>9</td>
<td>6</td>
<td>71,306.0</td>
</tr>
<tr>
<td>4</td>
<td>Massachusetts Institute of Technology</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>68,035.2</td>
</tr>
<tr>
<td>5</td>
<td>Columbia University</td>
<td>8</td>
<td>11</td>
<td>14</td>
<td>62,663.6</td>
</tr>
<tr>
<td>6</td>
<td>University of California, Berkeley</td>
<td>4</td>
<td>22</td>
<td>9</td>
<td>61,787.8</td>
</tr>
<tr>
<td>7</td>
<td>Yale University</td>
<td>11</td>
<td>7</td>
<td>11</td>
<td>60,686.7</td>
</tr>
<tr>
<td>8</td>
<td>University of Oxford</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>48,677.2</td>
</tr>
<tr>
<td>9</td>
<td>University of Chicago</td>
<td>9</td>
<td>8</td>
<td>10</td>
<td>47,178.7</td>
</tr>
<tr>
<td>10</td>
<td>Stanford University</td>
<td>2</td>
<td>15</td>
<td>2</td>
<td>45,926.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>University of Melbourne</td>
<td>57</td>
<td>36</td>
<td>28</td>
<td>11,962.1</td>
</tr>
<tr>
<td>53</td>
<td>University of Sydney</td>
<td>93</td>
<td>39</td>
<td>63</td>
<td>9,995.6</td>
</tr>
<tr>
<td>112</td>
<td>Australian National University</td>
<td>64</td>
<td>24</td>
<td>37</td>
<td>4,451.1</td>
</tr>
<tr>
<td>172</td>
<td>University of Queensland</td>
<td>90</td>
<td>46</td>
<td>65</td>
<td>2,772.0</td>
</tr>
</tbody>
</table>

The Rankings Cont.

Top 5 universities and the PIC obtained by each relation

<table>
<thead>
<tr>
<th>Relation</th>
<th>Harvard University</th>
<th>Princeton University</th>
<th>Massachusetts Institute of Technology</th>
<th>Columbia University</th>
<th>Stanford University</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:almaMater</td>
<td>114,387.1</td>
<td>68,121.6</td>
<td>65,404.4</td>
<td>48,694.0</td>
<td>39,707.7</td>
</tr>
<tr>
<td>dbo:education</td>
<td>9,745.1</td>
<td>2,535.4</td>
<td>1,682.5</td>
<td>10,484.6</td>
<td>4,652.5</td>
</tr>
<tr>
<td>dbo:employer</td>
<td>917.8</td>
<td>211.6</td>
<td>238.7</td>
<td>453.0</td>
<td>446.7</td>
</tr>
<tr>
<td>dbo:occupation</td>
<td>97.5</td>
<td>60.9</td>
<td>137.4</td>
<td>839.8</td>
<td>157.6</td>
</tr>
<tr>
<td>dbo:president</td>
<td>21.2</td>
<td></td>
<td></td>
<td></td>
<td>21.2</td>
</tr>
<tr>
<td>dbo:president</td>
<td>21.2</td>
<td></td>
<td></td>
<td></td>
<td>21.2</td>
</tr>
<tr>
<td>dbo:publisher</td>
<td>76.3</td>
<td>159.4</td>
<td>78.4</td>
<td>58.2</td>
<td>21.2</td>
</tr>
<tr>
<td>dbo:team</td>
<td>99.5</td>
<td>175.8</td>
<td></td>
<td>55.8</td>
<td>56.1</td>
</tr>
<tr>
<td>dbo:training</td>
<td>634.8</td>
<td>41.3</td>
<td>493.8</td>
<td>2,078.2</td>
<td>863.5</td>
</tr>
<tr>
<td>Total</td>
<td>125,979.3</td>
<td>71,306.0</td>
<td>68,035.2</td>
<td>62,663.6</td>
<td>45,926.4</td>
</tr>
</tbody>
</table>
Evaluation Results

Correlation of Scores

<table>
<thead>
<tr>
<th></th>
<th>PIC (Basic)</th>
<th>PIC</th>
<th>PIC (Basic)</th>
<th>PIC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pearson Correlation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SJTU</td>
<td>0.788</td>
<td>0.848</td>
<td>0.515</td>
<td>0.585</td>
</tr>
<tr>
<td>QS</td>
<td>0.553</td>
<td>0.68</td>
<td>0.439</td>
<td>0.643</td>
</tr>
<tr>
<td>THE</td>
<td>0.65</td>
<td>0.672</td>
<td>0.552</td>
<td>0.619</td>
</tr>
</tbody>
</table>
### Evaluation Results Cont.

#### Similarity with Other Systems

<table>
<thead>
<tr>
<th></th>
<th>SJTU</th>
<th>QS</th>
<th>THE</th>
<th>Average Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlap</td>
<td>0.61</td>
<td>0.51</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>PIC (Basic)</td>
<td>0.66</td>
<td>0.56</td>
<td>0.66</td>
<td>0.616</td>
</tr>
<tr>
<td>PIC</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td>0.628</td>
</tr>
<tr>
<td>Average</td>
<td>0.616</td>
<td>0.511</td>
<td>0.573</td>
<td>0.638</td>
</tr>
</tbody>
</table>

- PIC (Basic) and PIC are the two systems being compared.
- SJTU, QS, and THE represent different ranking systems.
- The table shows the overlap similarity between PIC (Basic), PIC, SJTU, QS, and THE, along with the average overlap.

### Diagram Description

- The graph visually represents the similarity between PIC (Basic) and PIC, SJTU, QS, and THE.
- The x-axis represents the similarity overlap, ranging from 0.40 to 0.70.
- The y-axis represents the different systems being compared.
- The bars indicate the similarity scores for each system.
### Pairwise Similarity of All Rankings (Average Overlap)

<table>
<thead>
<tr>
<th></th>
<th>PIC</th>
<th>SJTU</th>
<th>QS</th>
<th>THE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIC</td>
<td>1.0</td>
<td>0.669</td>
<td>0.628</td>
<td>0.638</td>
</tr>
<tr>
<td>SJTU</td>
<td>0.669</td>
<td>1.0</td>
<td>0.627</td>
<td>0.728</td>
</tr>
<tr>
<td>QS</td>
<td>0.628</td>
<td>0.627</td>
<td>1.0</td>
<td>0.721</td>
</tr>
<tr>
<td>THE</td>
<td>0.638</td>
<td>0.728</td>
<td>0.721</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Distribution of Information Content Regarding Top 500 Universities Across Continents and Countries

Percentage of Total

- Americas: Total PIC, Total Number of Top Universities
- Europe: Total PIC, Total Number of Top Universities
- Asia: Total PIC, Total Number of Top Universities
- Oceania: Total PIC, Total Number of Top Universities
- Africa: Total PIC, Total Number of Top Universities

- United States: Total PIC, Total Number of Top Universities
- England: Total PIC, Total Number of Top Universities
- Germany: Total PIC, Total Number of Top Universities
- Canada: Total PIC, Total Number of Top Universities
- Australia: Total PIC, Total Number of Top Universities
- Russia: Total PIC, Total Number of Top Universities
- France: Total PIC, Total Number of Top Universities
- Netherlands: Total PIC, Total Number of Top Universities
- Japan: Total PIC, Total Number of Top Universities
- Switzerland: Total PIC, Total Number of Top Universities
- Others: Total PIC, Total Number of Top Universities
Discussions

› High Similarity with SJTU Rankings
  - THE and QS incorporate subjective indicators (40% weight on survey)
  - SJTU is more objective (publications, awards, Fields Medal, etc.)

› PIC (Basic) vs. PIC –Based Rankings
  - Average of 8.5% difference
  - Still encouraging, with 51% to 62% similarity

› Pairwise High Similarity Between All Rankings
  - 60% to 75% Average Overlap

› Digital Divide Between American and universities in the rest of the world
  - Publish more on the (Semantic) Web
  - Contribute to Wikipedia
Conclusion and Future Work

› An information theory-based metric was developed for ranking using LOD
  - Further applications in information filtering, data visualization, multi-faceted browsing, and semantic navigation
  - Produces reasonable results with the extra advantage of low-cost data acquisition and replication.

› The need for a specific Linked University DB for university and research-related content.

› Future Work:
  - Rankings will be published on annual basis
  - A panel of academics will be asked to give the weights
  - Extract additional (and relevant) semantics from different parts of the Linked Open Data
Questions