Fusing Data with Correlations

Ravali Pochampally
UMass Amherst
Troo.jy

Anish Das Sarma
Google

Alexandra Meliou
AT&T Research

Divesh Srivastava
AT&T Research

### Knowledge base

<table>
<thead>
<tr>
<th>ID</th>
<th>KnowledgeTriple</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>{ Obama, profession, president}</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>t2</td>
<td>{ Obama, died, 1982}</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t3</td>
<td>{ Obama, profession, lawyer}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t4</td>
<td>{ Obama, religion, Christian}</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>t5</td>
<td>{ Obama, age, 50}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t6</td>
<td>{ Obama, support, White Sox}</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t7</td>
<td>{ Obama, administered by, John G. Roberts}</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>t8</td>
<td>{ Obama, surgical operation, 05/01/2011}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t9</td>
<td>{ Obama, profession, community organizer}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Correlations in web extraction

[Dong et al. PVLDB 2015]

- Correlations are richer than copying relationships
- There is an exponential number of correlation parameters, but we provide a scalable solution

### Synthetic datasets

- **Low precision**
  - Source quality: 0.2
  - Recall: 0.8
  - Precision: 0.6
  - F1: 0.8

- **High precision**
  - Source quality: 0.8
  - Recall: 0.2
  - Precision: 0.6
  - F1: 0.8

- **Low recall**
  - Source quality: 0.2
  - Recall: 0.8
  - Precision: 0.6
  - F1: 0.8

- **Bad to good quality**
  - Source quality: 0.8
  - Recall: 0.2
  - Precision: 0.6
  - F1: 0.8

### Real-world datasets

- **Restaurant:** 7 sources, 93 triples
  - Precision: 0.6
  - Recall: 0.8
  - F1: 0.8

- **Book:** 879 sources, 225 triples
  - Precision: 0.6
  - Recall: 0.8
  - F1: 0.8

- **ReVerb:** 6 extractors, 2407 triples
  - Precision: 0.6
  - Recall: 0.8
  - F1: 0.8

### Elastic approximation

Runtime comparison

<table>
<thead>
<tr>
<th>Source</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNION-25</td>
<td>0.39</td>
<td>0.56</td>
<td>0.46</td>
<td>124</td>
</tr>
<tr>
<td>UNION-50</td>
<td>0.14</td>
<td>0.32</td>
<td>0.31</td>
<td>3791</td>
</tr>
<tr>
<td>UNION-75</td>
<td>0.11</td>
<td>0.35</td>
<td>0.30</td>
<td>35</td>
</tr>
<tr>
<td>S-REVERB</td>
<td>0.7</td>
<td>0.66</td>
<td>0.70</td>
<td>30</td>
</tr>
<tr>
<td>LTM (10 iter)</td>
<td>0.49</td>
<td>0.53</td>
<td>0.51</td>
<td>35</td>
</tr>
<tr>
<td>PrecRec</td>
<td>2.6</td>
<td>0.3</td>
<td>0.35</td>
<td>2</td>
</tr>
<tr>
<td>PrecRecCorr</td>
<td>1.3</td>
<td>5.4</td>
<td>6</td>
<td>6746</td>
</tr>
<tr>
<td>PrecRecCorr-LTM</td>
<td>0.79</td>
<td>2.25</td>
<td>2452</td>
<td></td>
</tr>
</tbody>
</table>