A Collaborative Named Entity Focused URI Collection to Explore Web Archives

Workshop on Web Archiving and Digital Libraries

Sergej Wildemann & Helge Holzmann

June 6, 2019

L3S Research Center - University of Hannover - Germany
Introduction
Motivation

- Named entities are evolving
  - Cities grow
  - People change positions in their careers
- Information is spread on the Web
- Content is changed or deleted on web pages
- Web search engines forget or rerank resources
- Limited search in web archives by:
  - Exact URI
  - "Full-text"

Challenge

Accessing online resources that describe an entity over time
Example: Site Search in the Wayback Machine

- Indexed anchor texts
- Problem: Domains only, no specific paths or dates
Dataset Generation
Processing Pipeline

1. Entity collection
2. URI collection and assignment
3. URI unification
4. Ranking
5. Temporal enrichment
## Incorporated Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Entities</th>
<th>Classes</th>
<th>URIs</th>
<th>Tags</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>●</td>
<td></td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBpedia</td>
<td>□</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wikidata</td>
<td>□</td>
<td></td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delicious</td>
<td>□</td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>GWA</td>
<td>□</td>
<td></td>
<td>●</td>
<td></td>
<td>●</td>
</tr>
<tr>
<td>GWW</td>
<td>□</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
<tr>
<td>Wayback CDX</td>
<td>□</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
</tbody>
</table>

(●) extracted information – (□) used for joining
**Entity Collection**

- Potential entities: List of all Wikipedia article titles
- Filtering with DBpedia’s ontology:

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>1,500,000</td>
<td>22.7</td>
</tr>
<tr>
<td>Place</td>
<td>840,000</td>
<td>12.7</td>
</tr>
<tr>
<td>Creative Work</td>
<td>496,000</td>
<td>7.5</td>
</tr>
<tr>
<td>Organization</td>
<td>286,000</td>
<td>4.3</td>
</tr>
<tr>
<td>Other</td>
<td>2,378,000</td>
<td>36.0</td>
</tr>
<tr>
<td>N/A</td>
<td>1,100,000</td>
<td>16.7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>6,600,000</td>
<td>100</td>
</tr>
</tbody>
</table>
URI Collection

- **Wikidata** provides URIs directly or indirectly via identifiers

<table>
<thead>
<tr>
<th>official website</th>
<th><a href="https://barackobama.com/">https://barackobama.com/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 references</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reddit username</th>
<th>PresidentObama</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 reference</td>
</tr>
</tbody>
</table>

- **Wikipedia** articles contain "External Links" section
  - Official websites (usually without a path)
  - Databases like IMDb
- **GWA** and **GWW** were generated from the German Web Archive (1996-2013)
  - Searching for entities in anchor texts
  - GWW restricted the search to pages referenced from Wikipedia
  - Up to 10 URIs per entity
  - With most prominent years in dataset
Delicious contains tagged bookmarks with timestamps

<table>
<thead>
<tr>
<th>DATE</th>
<th>USERID</th>
<th>URI</th>
<th>TAG...</th>
</tr>
</thead>
</table>

- General idea: Matching of tags against entities
- Problem: Tags are single words
- Solution: Normalizing tags and entity titles
- Disambiguation terms must be found in tags
- Additional tags as metadata

**Normalization**

Entity: New York (State) ⇒ newyork
Tag: new-york ⇒ newyork
URI Unification

URI := protocol://domain[/path][?query][#fragment]

- Protocol removal
- Subdomain unification (www. prefix)
- Path stripping (index pages)
- Query parameter cleanup (empty or tracking keys)
- Fragment removal

Unification

http://www.example.org/index.html?foo=&ref=123#content
https://example.org/?ref=456#about
⇒ www.example.org
Ranking of Entity-URI Matches

- Provide initial votings in range $[1, 10]$ per dataset
- Wikipedia with links to homepages or databases
  - Voting: 10 (domains), 5 (URIs with path)
- Wikidata contains hand-picked URIs, but many indirect ones seem not useful
  - Voting: 10 (direct), 5 (indirect)
- GWA and GWW seem to have weak results
  - Voting: 3
- Delicious URI matches are ranked by relative number of users of a tag
  - Voting: 10 (most used tag) down to 1
Dataset by the Numbers

- 22.8 M URIs and 1.6 M described entities
  - 91.3% of entities have matching URIs
  - 13.4% of entities with at least one multi-source URI
  - 13.7 URIs per entity on average
- URIs for 70% of entities in both Wiki and GWW
- 35% in GWW and only 2.3% in Delicious

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Person</th>
<th>Org.</th>
<th>Place</th>
<th>C.W.</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>3.52</td>
<td>2.56</td>
<td>2.47</td>
<td>1.81</td>
<td>2.83</td>
</tr>
<tr>
<td>Wikidata</td>
<td>5.52</td>
<td>2.79</td>
<td>2.96</td>
<td>5.95</td>
<td>4.60</td>
</tr>
<tr>
<td>Delicious</td>
<td>22.15</td>
<td>51.43</td>
<td>63.53</td>
<td>61.43</td>
<td>55.86</td>
</tr>
<tr>
<td>GWA</td>
<td>6.54</td>
<td>7.16</td>
<td>6.52</td>
<td>8.49</td>
<td>7.14</td>
</tr>
<tr>
<td>GWW</td>
<td>4.43</td>
<td>5.32</td>
<td>4.63</td>
<td>6.10</td>
<td>5.13</td>
</tr>
</tbody>
</table>
Dataset Cross-Evaluation
Quality of Individual Datasets

Mean Reciprocal Rank

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{\infty} \frac{1}{\text{rank}_i}$$

- All URIs of a dataset as Q
- Avg. inverse rank in results
- Penalty for missing URIs

Results

- Delicious performs well overall
- URI selection of GWW better than GWA
Quality of Overlapping Results

nDCG – simplified in this environment:
- Only overlapping URIs
- Ideal rank: all results rank 1

Results
- Best URIs in Wiki
- GWA contains more of the popular URIs than GWW
Comparison with Web Search Engines

*Bing* results for 50,000 queries of our most promising entities

- 10 URIs from the first result page
- **Precision:**
  - Our top URI is in the result set of 83% of entities
  - Average Bing position: 2.26
- **Recall:**
  - 23.29% of all Bing URIs in our dataset
Conclusion
Conclusion

Ordered collection of annotated resources describing 1.6 M named entities over time.

- Combination of multiple diverse datasets
- Evaluation of dataset quality
- Promising results with respect to regular search engines

Future Work

- Integration of more data sources
  - Expansion covered time-frames and entities
  - Date tag evaluation
- Language awareness
- Improved entity matching
Explore Dataset Online

https://tempurion.l3s.uni-hannover.de