Collection Ranking & Selection for Federated Entity Search

Krisztian Balog, Robert Neumayer, and Kjetil Narvåg
Norwegian University of Science and Technology

Motivation

- Entities are ubiquitous
  Many information needs revolve around entities (people, products, organisations, places, etc.)
- Growing amount of structured data
  Again, organised around entities
- Entities are often searched by their name
  If we know (heard of) it, we just ask for it by the name

Top searches of 2011*

<table>
<thead>
<tr>
<th>Web search</th>
<th>Mobile search</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. iPhone</td>
<td>1. iPhone 5</td>
</tr>
<tr>
<td>2. Casey Anthony</td>
<td>2. Powerball</td>
</tr>
<tr>
<td>3. Kim Kardashian</td>
<td>3. MLB</td>
</tr>
<tr>
<td>4. Katie Perry</td>
<td>4. Scrabble cheat</td>
</tr>
<tr>
<td>5. Jennifer Lopez</td>
<td>5. Casey Anthony</td>
</tr>
<tr>
<td>8. Jennifer Aniston</td>
<td>8. Translator</td>
</tr>
</tbody>
</table>

* http://yearinreview.yahoo.com/

Top searches of 2011*

<table>
<thead>
<tr>
<th>Web search</th>
<th>Mobile search</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. iPhone</td>
<td>1. iPhone 5</td>
</tr>
</tbody>
</table>

"users have learned that search engine relevance decreases with longer queries and have grown accustomed to reducing their query (at least initially) to the name of an entity"

[Blanco et al., 2011]

* http://yearinreview.yahoo.com/

The Web of Data

![Graph showing the increase in Linking Open Data datasets from 2007 to 2011](image)

LOD in 2007

Ad-hoc entity search

At the 2010/11 Semantic Search Challenge

- Task
  Given a keyword query, targeting a particular entity, provide a ranked list of relevant entities (i.e., URIs)
- Queries
  Sampled from web search engine logs (142 in total)
- Data collection
  Billion Triple Challenge 2009 (BTC) dataset
- Relevance judgments
  On a 3-point scale, collected using crowdsourcing
In this talk

- Address the ad-hoc entity retrieval task in a distributed setting
  - The Web of Data is inherently distributed
  - Some data sources may not be crawlable at all
- Specifically, our focus is on the collection ranking and collection selection steps

Federated search

A typical broker-based architecture

1. Collection ranking
2. Collection selection
3. Result merging

Next: baseline models

for collection ranking and selection

1. Collection ranking
2. Collection selection
3. Result merging

Collection ranking

Collection-centric (CC)

- Lexicon-based method
  Treat and score each collection as if it was a single, large document

\[ P(c|q) \propto P(c) \cdot \prod_{t \in q} P(t|\exists c) \]

Collection selection

Top-K selection

- Choosing a fixed cutoff (K) ahead of time
  K is usually set between 5 and 20
**Our method: AENN**

"All that an Entity Needs is a Name"

- Exploit that entities are searched by their name
- The central broker maintains a complete dictionary of entity names (and corresponding identifiers)
- Utilise this information in the collection selection step to dynamically adjust the #collections selected

**AENN collection selection**

Balanced selection (AENN(b))

- Include collections from AENN until all from EC are covered

**AENN collection ranking**

Example

<table>
<thead>
<tr>
<th>CC</th>
<th>EC</th>
<th>AENN</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.15</td>
<td>0.65</td>
</tr>
<tr>
<td>B</td>
<td>0.3</td>
<td>0.6</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
<td>0.65</td>
</tr>
<tr>
<td>E</td>
<td>0.15</td>
<td>0.71</td>
</tr>
<tr>
<td>F</td>
<td>0.05</td>
<td>0.75</td>
</tr>
</tbody>
</table>

**AENN collection selection**

Precision-oriented selection (AENN(p))

- Only select collections returned by EC

**AENN collection selection**

Recall-oriented selection (AENN(r))

- Include collections from CC until all from EC are covered. This defines the cutoff point for AENN

**AENN collection selection**

Comparison of approaches

- Key observation
  Different methods—collection-centric (CC) vs. entity-centric (EC)—work best for different queries
- Idea
  Combination should give better results than any of the two methods alone

\[
AENN(c, q) = (1 - \lambda) \cdot CC(c, q) + \lambda \cdot EC(c, q)
\]
Experimental setup
Based on the 2010/11 Semantic Search Challenge

- **Distributed environment**
  Top 100 largest second-level domains from BTC
- **Relevance**
  Considered the relevant entities from each collection
- **Metrics**
  - Collection ranking: Standard IR metrics (MAP, MRR, nDCG)
  - Collection selection: Analogues of precision and recall, plus the avg. #coll. selected

Test collections

<table>
<thead>
<tr>
<th></th>
<th>BTC</th>
<th>BTC</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Entities</td>
<td>68.8M</td>
<td>60.5M</td>
<td>8.8M</td>
</tr>
<tr>
<td>#Collections</td>
<td>100</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>#Queries</td>
<td>136</td>
<td>116</td>
<td>130</td>
</tr>
<tr>
<td>Avg. #rel. entities/query</td>
<td>14.9</td>
<td>4.8</td>
<td>10.1</td>
</tr>
<tr>
<td>Avg. #rel. entities/collection</td>
<td>3.4</td>
<td>2.6</td>
<td>9.4</td>
</tr>
</tbody>
</table>

Results

Collection ranking (BTC)

![Graph showing MAP for name-only, full content, CC, EC, AENN]

Collection selection (BTC)

- **Precision**
- **Recall**
- **Avg. #coll. selected**

Collection selection (DBpedia)

- **Precision**
- **Recall**
- **Avg. #coll. selected**

Results

Different collection selection strategies (BTC\DBpedia)

- **Precision**
- **Recall**
- **Avg. #coll. selected**

Summary

- Addressed the task of ad-hoc entity retrieval in a distributed setting
- Introduced AENN, a novel collection ranking and selection method based on a lean name-based entity representation
- Showed experimentally that AENN can outperform standard baselines that consider all entity content
- Further, AENN can be geared towards high precision, high recall, or a balanced setting

Questions?

Contact
@krisztianbalog
krisztianbalog.com