Entity track 2010
Overview
Krisztian Balog, Pavel Serdyukov, Arjen P. de Vries
The Eighteenth Text Retrieval Conference (TREC 2010)
Gaithersburg, Maryland, USA, November 2010.

Overall aim

- Perform entity-oriented search tasks on the WWW
- Address information needs that are better answered by returning objects (instead of just any types of documents)

Example
Related Entity Finding

Input entity

Target type

Narrative

Name + Homepage

- Person
- Organization
- Product
- Location

Description of the relation in free text

Lessons learned from 2009

- Not that many entity pages in ClueWeb B
- Heavy use of Wikipedia/DBPedia

Key changes 2010

- ClueWeb English
- New entity type: location
- Single record submission format
- No supporting documents
- Wikipedia pages are not accepted
- Primary homepages are rewarded more

Entity homepage

Primary:
devoted to and in control of the entity

Relevant:
devoted to the entity, but not in control of the entity

Homepage

Primary:
devoted to and in control of the entity

Relevant:
devoted to the entity, but not in control of the entity

Wikipedia page of the entity
### Entity homepage

- **Homepage**
- **Primary:** devoted to and in control of the entity
- **Relevant:** devoted to the entity, but not in control of the entity
- **WikiPedia page of the entity**
- **News articles and blog posts, even if exclusively about the entity**

### Topics, assessment

- 20 topics from 2009
- Will be reassessed
- 50 new topics
- 47 of which are judged (pooled to depth 20)

### Qrels statistics

<table>
<thead>
<tr>
<th>Correct answers per topic</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homepages per entity</td>
<td>3.34</td>
</tr>
<tr>
<td>Primary</td>
<td>1.88</td>
</tr>
<tr>
<td>Relevant</td>
<td>1.46</td>
</tr>
</tbody>
</table>

### Results per topic

- **Query:**
  - 51
  - **Entity name:** National Institutes of Health (NIH)
  - **Entity URL:** clueweb09-en0009-58-39495
  - **Target entity:** organization
  - **Narrative:** What organizations comprise the National Institutes of Health (NIH)?

### Metrics

- **Main metric:** NDCG@R
- Gain=1 for relevant
- Gain=3 for primary
- RPrec and MAP for both relevance levels
Results per topic type

Number of topics

mean NDCG@R

Organization  Person  Location  Product

A typical approach

Input
(entity, target type, relation)

Set of candidate entities

Ranked list of entities

Homepages of entities

- Retrieving documents/snippets
- Query expansion
- Type filtering
- Deduplication
- Exploiting lists
- Heuristic rules
- Learning
-...

Entity List Completion

- Find entities that are engaged in a specific relation with an input entity
- Entities are represented by a unique URI in the Linked Open Data Cloud
- Topic definition includes example entities

Billion Triple Collection
(BTC-2009)

- Sample of the Linked Open Data cloud
- Semantic Web crawl
- ~1.14 billion statements
- 17GB gzipped

Results per submission

14 teams, 48 runs

mean NDCG@R

Topics

- 13 out of the 20 REF 2009 topics
- Excluded topics with complete result sets and topics without example entities
- Mapped type to the most specific class within the DBPedia ontology
- Mapped known relevant entity HPs to their URI in BTC-2009

Example topic
Assessments

- Done by track coordinators
- No relevant entities for 5 topics
- Gave up on judging one topic

Initial observations

- Many retrieved results are descriptors of the input entity itself
- Much more results of the correct type than in case of ClueWeb
- “Walk the graph” approaches fail to select on the correct relation

Results

(5 runs from 3 groups)

- later this week...

Questions?

http://ilps.science.uva.nl/trec-entity/