Finding Related Entities

Krisztian Balog
Norwegian University of Science and Technology

First International Workshop on Search and Mining Entity-Relationship Data (SMER’11), CIKM 2011
Glasgow, UK, Oct 2011

Overview of the TREC Entity track

Joint work with Arjen P. de Vries and Pavel Serdyukov

What is an entity?

An entity has...

• A unique identifier
• Name(s)
• Type
• Attributes
• Relations to other entities

Why entity retrieval?

“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”
(Salton, 1968)

“People want answers not documents”
(Unknown QA evangelist)

Entities are natural retrieval units

Research questions

• How to represent entities?
• How to represent information needs?
• How to match these representations?

Outline for this talk

• Background
• REF@TREC
• Future challenges
## Entity-oriented tasks

- Question Answering (@TREC)
- Expert Finding (@TREC)
- Entity Search (@INEX)

### Question Answering

- Direct answers to natural language questions
- Factual questions
  - Who invented the paper clip?
  - What’s the second tallest mountain in the world?
- List type questions
  - What women have worn Chanel clothing to award ceremonies?

### Typical QA architecture

1. Question analysis
2. Document retrieval
3. Passage retrieval
4. Answer extraction
5. Answer selection

### Evaluation

- Answer strings are evaluated
  - Correct, incorrect, not supported, not exact, ...
- Reusability of the test collection is problematic

### Expert Finding

- Find people who are experts on a given topic
  - TREC Enterprise track 2005-2008
  - People (entities) are represented by a unique ID or by their email addresses

### Task

#### Two principal approaches

1. Candidate (or profile-based) models
   - Create a textual model ("profile" document) for candidates from associated documents
   - Rank profiles using document retrieval techniques

2. Document models
   - Find documents that are relevant to the topic
   - Find out who is most strongly associated with these documents
**Model 1: Candidate models**

- Query
- Offline
- Online

**Model 2: Document models**

- Rank 1
- Rank n
- Online

---

**Entity Search in Wikipedia**

- Given a query, return a ranked list of entities
- INEX Entity ranking track 2007-2009
- Entities are represented by their Wikipedia page
- Topic definition includes target categories

---

**Example**

- Query: Movies with eight or more Academy Awards
- Entity: Titanic (1997 film)

---

**A possible approach**

\[
p(c|q) \propto p(q|e) \cdot p(c)
\]

\[
p(q|e) = \lambda \cdot p(T_q|T_e) + (1 - \lambda) \cdot p(C_q|C_e)
\]

**Term-based representation**

- Query model
- Entity model

**Category-based representation**

- Query model
- Entity model

---

**TREC Entity track**

2009-2011

- Building on previous evaluation efforts, expand in terms of **type** and **scale**
- Address information needs on the Web that are better answered by returning entities
Example entity-oriented information need

airlines that currently use Boeing 747 planes

Related Entity Finding

• Return a ranked list of entities (of a specified type) that engage in a given relationship with a given source entity

Why REF?

• Grounded in practical applications
• More than 50% of Web search queries target entities [Pound et al., WWW'10]

• Challenging
• Existing approaches are not directly applicable

• Meets pragmatic constraints
• Results sets are of reasonable size

Why REF? (2)

• An entity has...
• A unique identifier
• Name(s)
• Type
• Attributes
• Relations to other entities
Ingredients for an IR evaluation campaign

- Task
- Data collection
- Test topics
- Relevance assessments
- Evaluation metric(s)

TREC Entity 2009 Pilot

The Related Entity Finding (REF) task

- Given
  - Input entity, defined by its name and homepage
  - Target type (person, organization, or product)
  - Narrative (describing the nature of the relation in free text)
- Return
  - Ranked list of homepages of related entities

Information need:
airlines that currently use Boeing-747 planes

TREC topic definition:
<narrative>airlines that currently use Boeing-747 planes</narrative>
<entity_name>Boeing 747</entity_name>
<entity_URL>clueweb09-en0005-75-02292</entity_URL>

Example
7-HP1 Q0 clueweb09-en0003-03-28240 1 0.9874 KMR1PU
7-HP2 Q0 clueweb09-en0001-00-06792 1 0.9874 KMR1PU
7-HP3 Q0 clueweb09-en0000-71-05085 1 0.9874 KMR1PU
7-WP Q0 clueweb09-en0001-50-20366 1 0.9874 KMR1PU
7-NAME Q0 Asiana_Airlines 1 0.9874 KMR1PU

Required output

- Homepages of the entity (up to 3)
- Wikipedia page of the entity
- String answer that represents the entity
- Supporting documents (up to 10)
Data collection

- ClueWeb09 “Category B”
  - 50 million English webpages

Test topics

- Topic development is difficult
  - Not many entities have homepage in CatB
  - What is a homepage, anyway?

- Trivial topics

- Strategy
  - Hunt for entities on the Web, then map them back to ClueWeb
  - Burn rate: 1/5 - 1/10

Example topics

- Organizations that award Nobel prizes
- CDs released by the King’s Singers
- Companies that John Hennessey serves on the board of
- Universities that are members of the SEC conference for football
- Scotch whisky distilleries on the island of Islay

Relevance assessments

Run 1
Run 2
Run N

Results

<table>
<thead>
<tr>
<th>Name</th>
<th>NDCG@R</th>
<th>P@10</th>
<th>#rel</th>
<th>#pri</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purdue</td>
<td></td>
<td></td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>uogTr</td>
<td></td>
<td></td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>CAS</td>
<td></td>
<td></td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>NiCT</td>
<td></td>
<td></td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>UAms (Amst)</td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TUDelft</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BUPTPRIS</td>
<td></td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>UALR_CB</td>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>UIUC</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Waterloo</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EceUdel</td>
<td></td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>BIT</td>
<td></td>
<td>45</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>UAms (ISLA)</td>
<td></td>
<td>90</td>
<td>300</td>
<td>300</td>
</tr>
</tbody>
</table>

Variations on computing graded relevance

<table>
<thead>
<tr>
<th>Variations</th>
<th>Max gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP-only</td>
<td>2</td>
</tr>
<tr>
<td>HP+Name</td>
<td>3</td>
</tr>
<tr>
<td>WP-only</td>
<td>2</td>
</tr>
<tr>
<td>WP+WP</td>
<td>2</td>
</tr>
<tr>
<td>WP+WP+Name</td>
<td>3</td>
</tr>
</tbody>
</table>
Observations

- ClueWeb B dataset doesn’t contain enough entity pages
- Heavy use of Wikipedia/DBpedia
- Need more precise definitions of what constitutes an entity homepage
- Primary homepages are not rewarded enough
- Simplify entity record - traditional single-record TREC format is preferred

Changes in 2010

- ClueWeb English (500 million pages)
- New entity type: location
- Single record submission format
- No supporting documents
- Wikipedia pages are not accepted
- Primary homepages are rewarded more
Topics, assessment

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

Metrics

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

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

Results per submission

14 teams, 48 runs

Results per topic

\[
\text{query} \begin{align*}
\text{num} & : 51 \\
\text{entity_name} & : \text{National Institutes of Health (NIH)} \\
\text{entity_URL} & : \text{clueweb09-en0009-58-39495} \\
\text{target_entity} & : \text{organization} \\
\text{narrative} & : \text{What organizations comprise the National Institutes of Health (NIH)?}
\end{align*}
\]

\[
\text{query} \begin{align*}
\text{num} & : 51 \\
\text{entity_name} & : \text{Ford Motor Company} \\
\text{entity_URL} & : \text{clueweb09-en0120-17-13549} \\
\text{target_entity} & : \text{organization} \\
\text{narrative} & : \text{What companies build parts used in production of Ford vehicles?}
\end{align*}
\]
A variation:
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

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

```xml
<query>
<num>7</num>
<entity_name>Boeing 747</entity_name>
<entity_URL>clueweb09-en0005-75-02292</entity_URL>
<target_entity>organization</target_entity>
<narrative>Airlines that currently use Boeing 747 planes.</narrative>
<!-- Available only for the ELC task -->
<entity_URIs>
<URI>http://dbpedia.org/resource/Boeing_747</URI>
</entity_URIs>
<target_type_dbpedia>dbpedia-owl:Airline</target_type_dbpedia>
<examples>
<entity>
<URI>http://dbpedia.org/resource/Northwest_Airlines</URI>
</entity>
<entity>
<URI>http://dbpedia.org/resource/British_Airways</URI>
<URI>http://twitter.com/British_Airways</URI>
</entity>
...
</examples>
</query>
```

Observations

- Often difficult to judge URI relevance
- Many retrieved results are descriptors of the input entity itself
- Much more results of the correct type than in case of ClueWeb

TREC Entity 2011
Underway
A typical REF approach

- Input: entity, target type, relation
- Candidate entities
- Ranked list of entities
- Entity Homepages

  - Retrieving documents/snippets
  - Query expansion
  - ... (other steps)

  - Type filtering
  - Deduplication
  - Exploiting lists
  - Heuristic rules
  - Learning
  - ...

Summary

- Evaluation campaign on related entity finding on the Web
- Both the document web and the data web
- Interesting task, many different approaches
- End-to-end focus seems to be an issue
- Difficult to untangle the performance contributions of individual components

Future challenges

- More entity types and more fine-grained types
- Temporal aspects
- Result presentation
Questions?

@krisztianbalog
http://krisztianbalog.com