Multi-step Classification Approaches to Cumulative Citation Recommendation

Krisztian Balog
University of Stavanger

Naimdjon Takhirov, Heri Ramampiaro, Kjetil Norvåg
Norwegian University of Science and Technology

Motivation
- Maintain the accuracy and high quality of knowledge bases
- Develop automated methods to discover (and process) new information as it becomes available

Task
Cumulative Citation Recommendation
- Filter a time-ordered corpus for documents that are highly relevant to a predefined set of entities
- For each entity, provide a ranked list of documents based on their “citation-worthiness”

Collection and topics
- KBA stream corpus
  - Oct 2011 - Apr 2012
  - Split into training and testing periods
  - Three sources: news, social, linking
  - raw data 8.7TB
  - cleansed version 1.2TB (270G compressed)
  - stream documents uniquely identified by stream_id
- Test topics (“target entities”)
  - 29 entities from Wikipedia (27 persons, 2 org)
  - uniquely identified by urlname

Annotation matrix
<table>
<thead>
<tr>
<th>contains mention</th>
<th>non-relevant</th>
<th>relevant</th>
<th>central</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>G</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>no</td>
<td>garbage</td>
<td>neutral</td>
<td>relevant</td>
</tr>
</tbody>
</table>

Scoring

<table>
<thead>
<tr>
<th>Target entity: Aharon Barak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stream_id</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
<tr>
<td>1234567890</td>
</tr>
</tbody>
</table>

Cutoff
- Predefined for each target entity
- Example: 1000

TREC 2012 KBA track

Central?
Overview

- “Is this document central for this entity?”
- Binary classification task
- Multi-step approach
  - Classifying every document-entity pair is not feasible
  - First step do decide whether the document contains the entity
  - Subsequent step(s) to decide centrality

Components

- Mention detection
- Supervised learning

Identifying entity mentions

- Goals
  - High recall
  - Keep false positive rate low
  - Efficiency
- Detection based on known surface forms of the entity
  - urlname (i.e., Wikipedia title)
  - name variants from DBpedia
  - DBpedia-loose: only last names for people
- No disambiguation

Features

1. Document (5)
   - Length of document fields (body, title, anchors)
   - Type (news/social/linking)
2. Entity (1)
   - Number of related entities in KB
3. Document-entity (28)
   - Occurrences of entity in document
   - Number of related entity mentions
   - Similarity between doc and the entity’s WP page
4. Temporal (38)
   - Wikipedia pageviews
   - Average pageviews
   - Change in pageviews
   - Bursts
   - Mentions in document stream
     - Average volume
     - Change in volume
     - Bursts
Identifying entity mentions
Results on testing period

<table>
<thead>
<tr>
<th>Identification</th>
<th>Document-entity pairs</th>
<th>Recall</th>
<th>False positive rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>urname</td>
<td>41.2K</td>
<td>0.842</td>
<td>0.559</td>
</tr>
<tr>
<td>DBpedia</td>
<td>70.4K</td>
<td>0.974</td>
<td>0.701</td>
</tr>
<tr>
<td>DBpedia-loose</td>
<td>12.5M</td>
<td>0.994</td>
<td>0.998</td>
</tr>
</tbody>
</table>

---

**Summary**
- Cumulative Citation Recommendation task @TREC 2012 KBA
- Two multi-step classification approaches
- Four groups of features
- Differentiating between relevant and central is difficult

**Classification vs. Ranking**
[Balog & Ramampiaro, SIGIR’13]
- Approach CCR as a ranking task
- Learning-to-rank
  - Pointwise, pairwise, listwise
- Pointwise LTR outperforms classification approaches using the same set of features

---

Questions?

http://research.idi.ntnu.no/wislab/kbaaa