Cumulative Citation Recommendation: A Feature-aware Comparisons of Approaches

Gebrekirstos G. Gebremeskel, Jiyin He*, Arjen P. de Vries*, and Jimmy Lin@

*CWI, Amsterdam, The Netherlands
@University of Maryland, College Park

DEXA-TIR 2014, Munich, Germany
September 04, 2014
• IR
  - Machine learning
  - Uses many features

• Several studies
  - Compare IR approaches
  - Recommend best approach

• Scenario
  - Two methods (mtd 1 and mtd 2) and two feature sets (ftr 1 and ftr 2). Using ftr 1, mtd 1 outperforms mtd 2.
Question

- Does that mean \textit{mtd} 1 will still outperform \textit{mtd} 2 if we use \textit{ftr} 2 in place of \textit{ftr} 1?
Problem setting and dataset

• Problem setting:
  – Cumulative citation recommendation (CCR)
    • filtering a stream to identify those documents that are citation-worthy to Knowledge Base(KB) entities of interest

• Dataset:
  – TREC KBA-CCR-2012 dataset
    • 29 Wikipedia entities
    • Time-stamped stream of documents of news, social media content.
    • Relevance judgments for training and testing
State of the art CCR

• Three different approaches
  – String matching
  – Classification
    • Random Forest (CL-RF)
  – Learning to Rank (LTR)
    • Random forest (LTR-RF)

• LTR approaches outperform classification approaches for CCR task
A two-step approach

- Filtering using DBpedia name variants
- Subsequent classification or learning to rank
Feature selection

- Preliminary feature elimination
- Forward selection
## Best Scores

<table>
<thead>
<tr>
<th>Method</th>
<th>F</th>
<th>SU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC-RF</td>
<td>0.360</td>
<td>0.263</td>
</tr>
<tr>
<td>MC-LTR-RF</td>
<td>0.390</td>
<td>0.369</td>
</tr>
<tr>
<td>LRE-KBA</td>
<td>0.377</td>
<td>0.329</td>
</tr>
<tr>
<td>CL-RF</td>
<td>0.402</td>
<td>0.396</td>
</tr>
<tr>
<td>LTRE-RF</td>
<td>0.394</td>
<td>0.411</td>
</tr>
<tr>
<td>CL-J48</td>
<td>0.388</td>
<td>0.306</td>
</tr>
</tbody>
</table>
Summary

- Identified a few, but more powerful features
- Fair comparison of several approaches from previous studies
- Found out that classification approaches outperform learning-to-rank approaches
  - contrary to previous findings
Take home..

- Comparing approaches is problematic due to the interplay between the approaches themselves and the feature sets one chooses to use.