Effective Relevance Feedback for Entity Ranking

Tereza Iofciu, Gianluca Demartini, Nick Craswell, Arjen P. de Vries
Expert Finding in TREC-ENT (Enterprise Track)

Collection:
- Corpus: crawl of *.w3.org sites
- People: names of 1092 people who may be experts

Query:
- ‘information retrieval’

Results:
- A list of people who know about information retrieval
• Queries are lists of actors on the Web, e.g.
  – Query: 1930s
    • Answers: Fred Astaire, Charlie Chaplin, W.C. Fields, Errol Flynn, Clark Gable, Greta Garbo, etc
  – Query: action
    • Answers: Arnold Schwarzenegger, etc
• People
  – Expert Finding evaluation

• Actors
  – No evaluation initiative... yet?!

• Car companies, countries, museums, ...
  [i.e., insert your fav entity type here]

Entity Ranking!!!
Example INEX XER 2008 Topics

- Countries that have hosted FIFA Football World Cup tournaments: *countries; football world cup*

- Formula 1 drivers that won the Monaco Grand Prix: *racecar drivers; formula one drivers*

- Italian Nobel prize winners: *nobel laureates*

  ... 

Many examples on

Entity Ranking

- Topical query $Q$
- Entity (result) type $T_x$
- A list of entity instances $X_s$

- Systems employ XML element text, structure, links
Q

Title
Italian Nobel prize winners

Entities
Dario Fo (#176791)
Renato Dulbecco (#744909)
Carlo Rubbia (#44932)

Categories
Nobel laureates (#924)

Description
I want all the Italian people who won the Nobel prize.

Narrative
I want a list of people who were Nobel prize laureates in any field and have Italian nationality.
INEX XER Tasks

• Entity Ranking (ER)
  – Given Q and T, provide Xs
• List Completion (LC)
  – Given Q and Xs[1..m]
  – Return Xs[m+1..N]
• Entities (Xs) are represented as Wikipedia pages
• Binary relevance, MAP ($\text{xinfAP}^*$)

* A simple and efficient sampling method for estimating AP and NDCG.
Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. SIGIR'08
AND NOW...

HOW TO SOLVE ENTITY RANKING
Possible approaches to XER

- **Link structure** [Pehcevski et al. ECIR08]
- **Language Models** [Weerkamp et al. INEX08]
- **NLP based** [Demartini et al. LA-WEB08]
- **Ontology based** [Demartini et al. WISE08]
- **Passage retrieval** [Zaragoza et al. CIKM07]

- It is a recent task (2y): low effectiveness
- All previous work use categories
• Category structure
  – Unweighted entity-category relation
  – Missing categories
  – Noise (categories not expressing type)

• Our contribution:
  – Find good categories and improve results via Relevance Feedback
• User issues a query to a system

• Pseudo RFB
  – System uses top N retrieved entities to refine results

• Interactive RFB
  – User selects relevant entities in top K results
  – System uses relevant entities in top N to refine results
• Propagation of weights through a DAG for finding best categories

• Edges between article $i$ and category $j$
  – Hard edges: article $i$ belongs to category $j$
  – Soft Edges: article $i$ links to article $i'$ in category $j$
Relevance feedback for XER

- $h_j$: total incoming hard edges for category $j$
- $s_j$: total incoming soft edges for category $j$

\[
\text{catweight}_j = \frac{10^{h_j} + s_j}{\log(\text{catsize} + 50)}
\]

\[
\text{entityweight}_k = \left(\sum_{j=1}^{n} \text{catweight}_j\right) \times P_k
\]
Setting

- INEX XER 2008
  - 600k Wikipedia articles
  - 35 topics
  - 32 Runs used as baselines
- Seed for our algorithm
  - Example entities
  - Pseudo RFB
  - Interactive RFB
- Fusion with baseline
Experimental Results

- Fusion with baseline (Example entities)

\[ \text{score}(e, q) := \lambda \cdot \text{baseline}(e, q) + (1 - \lambda) \cdot \text{LinkBased}(e, q) \]
Experimental Results

- Fusion with baseline (Pseudo RFB)
## Experimental Results

Table 5: Average absolute improvement of xinfAP for different values of $k$ in the pseudo relevance feedback and in the interactive relevance feedback cases.

<table>
<thead>
<tr>
<th></th>
<th>K=5</th>
<th>K=10</th>
<th>K=15</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudo RFB</td>
<td>0.050</td>
<td>0.057</td>
<td>0.051</td>
<td>0.040</td>
</tr>
<tr>
<td>Interactive RFB</td>
<td>0.083</td>
<td>0.103</td>
<td>0.112</td>
<td>0.118</td>
</tr>
<tr>
<td>AVG relevant in top-$k$</td>
<td>1.92</td>
<td>3.57</td>
<td>4.88</td>
<td>6.04</td>
</tr>
</tbody>
</table>

![Graph showing the relationship between original xinfAP and k values.]
Experimental Results

- Early Precision

Table 6: Expected P@20 measured for different values of $k$ in the interactive relevance feedback case.

<table>
<thead>
<tr>
<th></th>
<th>K=5</th>
<th>K=10</th>
<th>K=15</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.204</td>
<td>0.204</td>
<td>0.204</td>
<td>0.204</td>
</tr>
<tr>
<td>combination $\lambda = 0.5$</td>
<td>0.222</td>
<td>0.225</td>
<td>0.228</td>
<td>0.247</td>
</tr>
<tr>
<td>Relevance Feedback</td>
<td>0.265</td>
<td>0.289</td>
<td>0.289</td>
<td>0.29</td>
</tr>
</tbody>
</table>
Unique contributions

- **Pseudo RFB**

<table>
<thead>
<tr>
<th></th>
<th>K=5</th>
<th>K=10</th>
<th>K=15</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>relevant in baseline</td>
<td>5.746</td>
<td>4.820</td>
<td>4.770</td>
<td>4.752</td>
</tr>
<tr>
<td>relevant in pseudo RF</td>
<td>3.932</td>
<td>4.357</td>
<td>4.393</td>
<td>4.227</td>
</tr>
<tr>
<td>relevant in both</td>
<td>12.146</td>
<td>13.073</td>
<td>13.123</td>
<td>13.141</td>
</tr>
<tr>
<td>missed relevant</td>
<td>4.489</td>
<td>4.064</td>
<td>4.029</td>
<td>4.195</td>
</tr>
</tbody>
</table>

- **Interactive RFB**

<table>
<thead>
<tr>
<th></th>
<th>K=5</th>
<th>K=10</th>
<th>K=15</th>
<th>K=20</th>
</tr>
</thead>
<tbody>
<tr>
<td>relevant in baseline</td>
<td>8.191</td>
<td>6.545</td>
<td>5.941</td>
<td>5.470</td>
</tr>
<tr>
<td>relevant in pseudo RF</td>
<td>2.450</td>
<td>3.114</td>
<td>3.496</td>
<td>3.677</td>
</tr>
<tr>
<td>relevant in both</td>
<td>9.702</td>
<td>11.348</td>
<td>11.952</td>
<td>12.423</td>
</tr>
<tr>
<td>missed relevant</td>
<td>5.971</td>
<td>5.307</td>
<td>4.925</td>
<td>4.745</td>
</tr>
</tbody>
</table>
Per topic analysis

- RFB for doc search
  - works well on average
  - on some queries performs badly

- 15 topics over 12 systems had 0 relevant after RFB
  - All system having low AvgPrec: not a good seed

- 3 topics over 6 systems having 0 AvgPrecs had 0.16 after RFB
  - Exceptional case: 40 relevant
Conclusions

- Using top 10 results yields to best results
- Interactive RFB yields to best results
- RFB for XER is beneficial
  - differently from doc search
  - for all retrieval methods

- Limitations
  - Single test collection
  - Non-optimal parameters for the model
Thanks