Entity Oriented Search (EoS)

- Popular Approaches
- MRF for information retrieval
- MRF for Entity Oriented search
- Entity Document Scoring
- Entity Type Scoring
- Entity Name Scoring

Evaluation

- Benchmarks: INEX entity tracks 2007-2009
- Experimental Results
- Summary and future work

EoS: Profile based Approach (Craswell et al 2001):

- Represent each entity by a virtual document (a profile)
- e.g.
  - Entity home-page
  - Concatenating passages mentioning the entity
  - Rank those profiles according to their relevance to the query
  - Using standard IR ranking techniques
- Difficulties:
  - Co-resolution and name disambiguation
  - Profiling is not an easy task
EoS: Voting approach (Balog06, MacDonald09)

- Any relevant document is a “voter” for the entity mentioned within its content.

\[ Score(p,q) = \sum_d \text{Score}(d,q) \times \text{Score}(p,d) \]

- What is the ratio behind?
  - An entity mentioned many times in relevant (top retrieved) docs is more likely to be relevant on the given topic.

Markov Random Fields for IR (Metzler & Croft 2005)

- Full Independence
  \[ P(D\mid D) \]

- Sequential dependence
  \[ C_1(q\mid D) \]

- Full dependence
  \[ C_0(#wN(q\mid D)) \]

MRF for EoS

- Entity Type
  \[ P(E\mid Q) \]

- Entity Name
  \[ P(E\mid Q) \]

- Entity Document
  \[ P(E\mid Q) \]

MRF based Entity Document Scoring

- We consider cliques of the three types:
  - Full Independent (C\(_I\))
  - Sequential dependent (C\(_S\))
  - Full dependent (C\(_D\))

- The feature functions \(f_I(q)\) over clique of type \(l\) (\(l\) in \(T,O,U\))
  - \(f_I(q)\) measures how well the clique’s terms represent the entity document.
  - Based on Dirichlet smoothed language model:
    \[ f_I(q) = \log \left( \frac{\text{Freq}(q,E_t)}{\text{Freq}(q,T)} \right) \]

- For \(C_D\) and \(C_U\) we replace \(q\) with \(q_{\text{top}}\) and \(\#wN(q_{\text{top}})\) respectively.

- The entity document scoring function aggregates the feature functions over all clique types:
  \[ P(E\mid Q) \]

Entity type Scoring

- \(f_T(c)\) is defined over a single clique composed of \(E_t\) and \(Q_T\)
  \[ P(E_t\mid Q_T) \]

- \(d(q_t;E_T)\) - the type distance, is domain dependent
- In our experiments we measured the distance in the Wikipedia category graph
  - The minimal path length between all pairs of the query and the entity’s page categories

Entity Name Scoring

- Clique types:
  - Query terms independent – SE\(_N\) - a single node clique containing the entity name alone
    - Equivalent to the voting approach

- Query terms dependent – Consider proximity with the query terms
  - Does the entity is usually mentioned in proximity to query terms
    - In analogy to document scoring
      - \(T_{\text{top}}\) Full independent
      - \(O_{D}\), Sequential dependent
      - \(O_{U}\), Full dependent
Entity Name Scoring \( P(E\mid Q) \) (cont.)

- Local approach
  - Measure the relationship (e.g., proximity) between the query terms and the entity name in the top retrieved documents
- Global approach
  - Measure the PMI between the query term(s) and the entity name in the whole collection
  - PMI – the pointwise mutual information – the likelihood of finding one term in proximity to another term

\[
P(E\mid Q) = \sum_{A\in\text{local}} \sum_{n+k} f_n^k (c) \\
A = \{S, T, O, U, PMI_1, PMI_2, PMI_3\}
\]

Parameter tuning (cont) – Coordinate Ascent

- The parameters of the scoring function were tuned using the Coordinate Ascent algorithm for each benchmark
- Optimization process was done separately for each dataset, using the training topics
- Performance was estimated using the test topics

<table>
<thead>
<tr>
<th>Entity Property</th>
<th>Symbol</th>
<th>Parameter name</th>
<th>Optimal Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TD</td>
<td>N</td>
<td>Query term proximity window size</td>
<td>10</td>
</tr>
<tr>
<td>ET</td>
<td>( g_{\text{max}} )</td>
<td>Max. distance in category graph</td>
<td>5</td>
</tr>
<tr>
<td>alpha</td>
<td></td>
<td>Category score decay</td>
<td>3</td>
</tr>
<tr>
<td>EN</td>
<td>R</td>
<td>Top-docs for voting score computation</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>Entity name terms proximity window</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>( \lambda )</td>
<td>Top-docs for entity expansion</td>
<td>10</td>
</tr>
</tbody>
</table>

References:

1. MRF for ExS: JIWES 2012, Portland OR
2. MRF for ExS: JIWES 2012, Portland OR
3. MRF for ExS: JIWES 2012, Portland OR
4. MRF for ExS: JIWES 2012, Portland OR
5. MRF for ExS: JIWES 2012, Portland OR
6. MRF for ExS: JIWES 2012, Portland OR
7. MRF for ExS: JIWES 2012, Portland OR
8. MRF for ExS: JIWES 2012, Portland OR
9. MRF for ExS: JIWES 2012, Portland OR
10. MRF for ExS: JIWES 2012, Portland OR
11. MRF for ExS: JIWES 2012, Portland OR
12. MRF for ExS: JIWES 2012, Portland OR
13. MRF for ExS: JIWES 2012, Portland OR
14. MRF for ExS: JIWES 2012, Portland OR
15. MRF for ExS: JIWES 2012, Portland OR
16. MRF for ExS: JIWES 2012, Portland OR
17. MRF for ExS: JIWES 2012, Portland OR
18. MRF for ExS: JIWES 2012, Portland OR
19. MRF for ExS: JIWES 2012, Portland OR
20. MRF for ExS: JIWES 2012, Portland OR

INEX – Entity ranking track 2007-2009

- Entity Ranking (XER)
  - Return entities that satisfy a topic described in natural language text
  - List Completion (LC)
    - Complete a partial list of given answers
  - Entities:
    - Must have a Wikipedia page
    - Each answer should contain an article about mammal which can be a part of any circus show
    - Entities:
      - "mammals" category
        - "asiatic", "african", "indian"
      - "mammals" category
        - " despre" alternative
      - "carnivore", "omnivore"

Evaluation

The INEX Entity Ranking track 2007-2009

- The Collection – Wikipedia articles
- A retrievable entity must have a Wikipedia page
- No need for a third party named-entity extraction tool
- Each entity has a unique name, document and type (WP categories)
- INEX topics perfectly fit our model looking for relevant entities to a given topic
- Metrics: MAP for 2007, iMAP for 2008-2009

<table>
<thead>
<tr>
<th>Data set</th>
<th>Wikipedia Year</th>
<th>#docs</th>
<th>Train topics</th>
<th>Test topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>2006</td>
<td>659,388</td>
<td>28</td>
<td>46</td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td>74</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>2009</td>
<td>2,666,190</td>
<td>-</td>
<td>55</td>
</tr>
</tbody>
</table>

INEX topics perfectly fit our model looking for relevant entities to a given topic

- Evaluation
  - The INEX Entity Ranking track 2007-2009
  - The Collection – Wikipedia articles
  - A retrievable entity must have a Wikipedia page
  - No need for a third party named-entity extraction tool
  - Each entity has a unique name, document and type (WP categories)
  - INEX topics perfectly fit our model looking for relevant entities to a given topic
  - Metrics: MAP for 2007, iMAP for 2008-2009
Summary

In this work we presented an entity ranking model using the MRF framework which integrates
- Profile approach: query E-document relationship
- Voting approach: query E-name relationship
- Type filtering approach: query E-type relationship

Experiments over INEX benchmarks showed that
- Performance is relatively high and comparative to leading INEX systems
- Using dependence models did not result in significant improvement over Full Independence model.
- Global based name scoring outperforms local based name scoring

Future work
- Explore this model with additional data collections, specifically, large web collections
- Using additional entity properties, e.g., exploring the entity graph
- Further investigation of the dependence models

Results improved significantly when type and name scoring were added
Final Results are superior to top INEX 2007,2008, and comparable to 2009
Dependence models (SD, FD) have not improved over Independence model (FI) ???
Global based name scoring (PMI) outperforms local based name scoring

Thank You!

Questions?