ENTITY RANKING USING WIKIPEDIA AS A PIVOT

Authors: Rianne Kaptein, Pavel Serdyoukov, Arjen de Vries, Jaap Kamps
Keywords: Web Entity Ranking, Wikipedia

Presentation: Ki-Hyoun Kim
Webscience Course Winter Term 2010/2011, 25.01.2011
Overview

2. INTRODUCTION

3. FROM WIKIPEDIA ENTITIES TO WEB ENTITIES AND BACK

4. ENTITY RANKING ON WIKIPEDIA

5. ENTITY RANKING ON THE WEB

6. CONCLUSIONS
Introduction: Entity Ranking

Entity Ranking

- wikipedia.org/Apple
- www.apple.com/macbookpro
- www.apple.com/imac

Document retrieval

DOC1
Introduction: Entity Ranking

Differences between entity ranking and document retrieval:

1. returned documents have to represent an entity

2. this entity belongs to the chosen entity type

3. return the entity (the document representing the entity) only once
Introduction: Entity Ranking

**Goal**: return a list of relevant entities for a query

*Example*

*Query*: „Countries where I can pay in Euro“

*Results*: germany, spain, italy, ...etc.
Introduction: Entity Ranking

Possible ways to rank web entities:

- use structure of the web (links, anchor text)
- use wikipedia
Introduction: Entity Ranking

- More than 1 page representing the same entity

- Entity Ranking: return the most relevant page for each entity

Wikipedia is used as follows:

- Entities = Wikipedia Pages
- Name of entity = Title of Page
- Content of Page = Representation of the Entity

Pages can be associated with Wikipedia categories: type, topical, administrative

Administrative:
„Pages to be revised“ (for administrative purpose)

Topical:
„Barack Obama“ (topical relevant pages)

Type:
„People from Westminster“ , „Museums in Michigan“ (entity type)

- Advantages:
  - No problems with duplicates
  - Each entity is represented only once in Wikipedia

Entity retrieval:

1. Associate target entity types with the query (section 4.1)
2. Rank Wikipedia pages according to their similarity with the query and target entity types
3. Find web entities corresponding to the Wikipedia entities
Introduction: Research Questions

- What is the range of entity ranking topics, which can be answered using Wikipedia?
- When we find relevant Wikipedia entities, can we find the relevant web entities that correspond to the Wikipedia entities?
- Can we exploit category information to improve entity ranking queries?
- Can we automatically assign entity types to natural language queries?
- Can we improve web entity ranking by using Wikipedia as a pivot?
- Can we automatically enrich Wikipedia with additional links to homepages of found entities?
Related Works

- Entity search engines
  - NAGA – semantic search engine using a knowledge base
  - ESTER – combining full-text on Wikipedia with ontology search in YAGO
- INEX
- TREC
3. FROM WIKIPEDIA ENTITIES TO WEB ENTITIES AND BACK

3.1 From Web to Wikipedia

- Wikipedia has 3,147,000 Articles after 9 years
- Notability impeding Wikipedia's growth: Editors decision
3. FROM WIKIPEDIA ENTITIES TO WEB ENTITIES AND BACK

3.1 From Web to Wikipedia

- Results of analyzing the list of relevant entities of 20 queries:

<table>
<thead>
<tr>
<th>Table 1: Topic and Entity Coverage in Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td># Topics</td>
</tr>
<tr>
<td>- with entities in Wikipedia</td>
</tr>
<tr>
<td># Entities</td>
</tr>
<tr>
<td>- with Wikipedia pages</td>
</tr>
</tbody>
</table>
3. FROM WIKIPEDIA ENTITIES TO WEB ENTITIES AND BACK

3.1 From Web to Wikipedia

- Results of analyzing the list of relevant entities of 20 queries:

<table>
<thead>
<tr>
<th>Table 1: Topic and Entity Coverage in Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td># Topics</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>- with entities in Wikipedia</td>
</tr>
<tr>
<td>17 (85%)</td>
</tr>
<tr>
<td># Entities</td>
</tr>
<tr>
<td>198</td>
</tr>
<tr>
<td>- with Wikipedia pages</td>
</tr>
<tr>
<td>160 (81%)</td>
</tr>
</tbody>
</table>
3. FROM WIKIPEDIA ENTITIES TO WEB ENTITIES AND BACK

3.2 From Wikipedia to Web

- Can we use Wikipedia to search for web entities?
- Wikipedia has External links section
- Official guidelines: Articles about organisations, people, website should link to the official site, by convention listet first
4. ENTITY RANKING ON WIKIPEDIA

- Can we use category information to improve entity ranking queries?
- Can we automatically assign entity types to natural language queries?
4.1 Entity Types

- 4.1.1 Entity Type Assignment
  - Pseudo-Relevance Feedback using categories
  - Assign most frequent categories instead of most frequent terms of the top ranked results
  - Exclude categories that occur only once
4.1 Entity Types

• 4.1.1 Entity Type Assignment
  – Pseudo-Relevance Feedback using categories
4.1 Entity Types

• 4.1.1 Entity Type Assignment
  – Pseudo-Relevance Feedback using categories
4.1 Entity Types

• 4.1.2 Scoring Entities
  – computing the likelihood of the query terms occurring in the document
  – Using a language model with Jelinek-Mercer smoothing and uniformly distributed prior document probabilities:

\[ P(q_1, \ldots, q_n | d) = \sum_{i=1}^{n} \lambda P(q_i | d) + (1 - \lambda) P(q_i | D) \]

\( D \): Wikipedia Document collection

\( q_1, \ldots, q_n \): Query terms

\( d \): document
4.1 Entity Types

• 4.1.2 Scoring Entities
  - Maximum likelihood estimation of the probability for a term occurring in a category name:

\[
P(t_1, \ldots, t_n | C) = \sum_{i=1}^{n} \lambda P(t_i | C) + (1 - \lambda) P(t_i | D)
\]

\(t_1, \ldots, t_n\): Category terms
\(C\): Category Name
\(D\): Wikipedia Document collection (used for smoothing, avoiding division by zero)
4.1 Entity Types

• 4.1.2 Scoring Entities
  - Calculating the similarity between two categories using KL-divergence:

  \[ S_{cat}(C_t|C_d) = -D_{KL}(C_t|C_d) \]  
  \[ = -\sum_{t \in C_t} \left( P(t|C_t) \times \log \left( \frac{P(t|C_t)}{P(t|C_d)} \right) \right) \]

  \[ d: \text{a document, i.e. an answer entity} \]
  \[ C_t: \text{target category} \]
  \[ C_d: \text{category assigned to a document} \]
4.1 Entity Types

- 4.1.2 Scoring Entities

Entity type score:

\[ S_{cat}(d|QT) = \arg\max_{C_t \in QT} \arg\max_{C_d \in d} S_{cat}(C_t|C_d) \]  

Entity type score for document \( d \) in relation to a query topic \( QT \)

Maximum of the scores of all target document categories
4.1 Entity Types

4.1.2 Scoring Entities

- The final score:

\[ S(d|QT) = \mu P(q|d) + (1 - \mu) S_{cat}(d|QT) \] (6)

- Final score

- Linear combination of normalized scores
4.2 Experimental Setup

- Generic approach for entity types
- TREC uses 3 entity types:
  - people, organisations, products
- Advantages
  - Clear,
  - Few options,
  - easily selectable by users
- Disadvantages
  - cover only a small range of queries
- Manually assigned more specific types to TREC ranking topics
4.2 Experimental Setup

- Rerank the 2500 Results of the baseline run
  - Manually assigned: assigned manually by the authors
  - Automatically assigned: assigned by pseudo-relevance-feedback
- P10, NDCG for TREC
- P10, MAP for INEX
4.3 Experimental Results

Table 3: Wikipedia retrieval results on TREC topics

<table>
<thead>
<tr>
<th>Cats</th>
<th>μ</th>
<th>#Rel</th>
<th>P10</th>
<th>NDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1</td>
<td>78</td>
<td>0.1200</td>
<td>0.0797</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.7</td>
<td>74⁻</td>
<td>0.1500⁻</td>
<td>0.0980⁻</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.8</td>
<td>75⁻</td>
<td>0.1500⁻</td>
<td>0.0971⁻</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.9</td>
<td>79⁻</td>
<td>0.1500⁻</td>
<td>0.0969⁻</td>
</tr>
<tr>
<td>Man.</td>
<td>0.4</td>
<td>89⁻</td>
<td><strong>0.1750⁻</strong></td>
<td>0.1123⁰</td>
</tr>
<tr>
<td>Man.</td>
<td>0.5</td>
<td>91⁻</td>
<td><strong>0.1750⁻</strong></td>
<td><strong>0.1193⁰</strong></td>
</tr>
<tr>
<td>Man.</td>
<td>0.8</td>
<td><strong>96⁰</strong></td>
<td>0.1700⁰</td>
<td>0.1132⁰</td>
</tr>
</tbody>
</table>

Significance of increase or decrease over baseline according to t-test, one-tailed, at significance levels 0.05⁰, 0.01⁰, and 0.001⁰.

The results are expressed in the number of retrieved relevant pages.
4.3 Experimental Results

<table>
<thead>
<tr>
<th>Cats</th>
<th>μ</th>
<th>#Rel</th>
<th>P10</th>
<th>NDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1</td>
<td>78</td>
<td>0.1200</td>
<td>0.0797</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.7</td>
<td>74*</td>
<td>0.1500*</td>
<td>0.0980*</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.8</td>
<td>75*</td>
<td>0.1500*</td>
<td>0.0971*</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.9</td>
<td>79*</td>
<td>0.1500*</td>
<td>0.0969*</td>
</tr>
<tr>
<td>Man.</td>
<td>0.4</td>
<td>89*</td>
<td>0.1750*</td>
<td>0.1123*</td>
</tr>
<tr>
<td>Man.</td>
<td>0.5</td>
<td>91*</td>
<td>0.1750*</td>
<td>0.1193*</td>
</tr>
<tr>
<td>Man.</td>
<td>0.8</td>
<td>96*</td>
<td>0.1700*</td>
<td>0.1132*</td>
</tr>
</tbody>
</table>

Significance of increase or decrease over baseline according to t-test, one-tailed, at significance levels 0.05(°), 0.01(*), and 0.001(**).
### 4.3 Experimental Results

Table 4: Wikipedia retrieval results on INEX 2006-2008 topics

<table>
<thead>
<tr>
<th>Cats</th>
<th>$\mu$</th>
<th>#Rel</th>
<th>P10</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1</td>
<td>1142</td>
<td>0.2405</td>
<td>0.1948</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.7</td>
<td>1239°</td>
<td>0.2949°</td>
<td>0.2602°</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.8</td>
<td>1279°</td>
<td>0.2987°</td>
<td>0.2686°</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.9</td>
<td>1289°</td>
<td>0.2937°</td>
<td>0.2561°</td>
</tr>
<tr>
<td>Man.</td>
<td>0.7</td>
<td>1346°</td>
<td>0.3797°</td>
<td>0.3245°</td>
</tr>
<tr>
<td>Man.</td>
<td>0.8</td>
<td>1361°</td>
<td>0.3620°</td>
<td>0.3048°</td>
</tr>
<tr>
<td>Man.</td>
<td>0.9</td>
<td>1327°</td>
<td>0.3241°</td>
<td>0.2711°</td>
</tr>
</tbody>
</table>
4.3 Experimental Results

<table>
<thead>
<tr>
<th>Cats</th>
<th>$\mu$</th>
<th>#Rel</th>
<th>P10</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1</td>
<td>1042</td>
<td>0.2164</td>
<td>0.1674</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.8</td>
<td>911°</td>
<td>0.2382</td>
<td>0.1993°</td>
</tr>
<tr>
<td>Auto.</td>
<td>0.9</td>
<td>982°</td>
<td>0.2509</td>
<td>0.2014°</td>
</tr>
<tr>
<td>Man.</td>
<td>0.6</td>
<td>1171°</td>
<td>0.3145°</td>
<td>0.2376°</td>
</tr>
<tr>
<td>Man.</td>
<td>0.7</td>
<td>1178°</td>
<td>0.3127°</td>
<td>0.2396°</td>
</tr>
<tr>
<td>Man.</td>
<td>0.9</td>
<td>1180°</td>
<td>0.2982°</td>
<td>0.2350°</td>
</tr>
</tbody>
</table>
5. ENTITY RANKING ON THE WEB

- Can we improve web entity search by using Wikipedia as a pivot?
- Can we automatically enrich Wikipedia with additional external Links to homepages of entities?
5.1 Experimental results

<table>
<thead>
<tr>
<th>Run</th>
<th>Full Text</th>
<th>Wikipedia Link Cat+Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel. WP</td>
<td>73</td>
<td>73 - 57°</td>
</tr>
<tr>
<td>Rel. HP</td>
<td>244</td>
<td>69° 70°</td>
</tr>
<tr>
<td>Rel. All</td>
<td>316</td>
<td>134° 121°</td>
</tr>
</tbody>
</table>

| NDCG Rel. WP | 0.2119    | 0.2119 - 0.1959°        |
| NDCG Rel. HP | 0.1919    | 0.0820° 0.0830°         |
| NDCG Rel. All| 0.2394    | 0.1429° 0.1542°         |

| Primary WP   | 78        | 78 - 96°                |
| Primary HP   | 6         | 29° 34°                 |
| Primary All  | 86        | 107° 130°               |

| P10 pr. WP   | 0.1200    | 0.1200 - 0.1700°        |
| P10 pr. HP   | 0.0050    | 0.0300° 0.0400°         |
| P10 pr. All  | 0.1200    | 0.1300° 0.1850°         |

| NDCG pr. WP  | 0.1184    | 0.1184 - 0.1604°        |
| NDCG pr. HP  | 0.0080    | 0.0292° 0.0445°         |
| NDCG pr. All | 0.1041    | 0.1292° 0.1610°         |

Full text retrieval works well

Doesn't work for finding pr. Wikipedia pages (fails for pr. HP)
6. CONCLUSION

- Standard Text retrieval finds relevant Web Pages, but not primary homepages of entities
- Wikipedia as a pivot performs well in finding primary homepages
- Combination of external links / wikipedia anchor text improves entity retrieval
- Approach is based on three assumptions:
  - Entity coverage in Wikipedia is large enough
  - We are able to find entities in Wikipedia
  - We can map Wikipedia entities to the primary homepages
6. CONCLUSION

• Exploiting the structured information like Wikipedias structured pages improves search in unstructured resources like the web
Q?