Evaluation of an Information Retrieval System for the Semantic Desktop using Standard Measures from Information Retrieval

aposdle – New ways ...
... to work, learn and collaborate!

Peter Scheir, Michael Granitzer, Stefanie N. Lindstaedt
Starting point

- We built an information retrieval system for the Semantic Desktop [Scheir et al. 2007]
  - Semantic Web technology in desktop environment

- How evaluate it?
Current situation

- System for information retrieval in the Semantic Web / Desktop exist (see talk on Wednesday! ;)
- Not many are evaluated using classical measures
- No standard test corpus is available
Our work

- Evaluation of an information retrieval system for the semantic desktop

- Using:
  - Precision at ranks 10, 20 and 30
  - Inferred average precession (infAP)
Basic search approach

- Retrieval of documents based on concepts stemming from a knowledge representation
- Information retrieval, list ranked by relevance
- Network model
Extended search approach

- Query and result expansion
- Relate similar concepts by semantic similarity
- Relate similar documents by textual similarity
- Model associations in network and consider them during search
- See if approach provides better results as without associations
Our corpus

- Knowledge and document base from first release of APOSDLE system
- Requirements Engineering as application domain
  - Requirements Engineering ontology
  - Requirements Engineering document base
- Statistics
  - 70 concepts in ontology
  - 1016 documents in document base
  - 21 concepts used for annotation
  - 496 documents annotated
Evaluation approach

- **Configurations**
  - 8 different system configurations (with semantic and/or textual similarity)

- **Queries**
  - 79 queries

- **Judgments**
  - First 30 results
  - Results overlapping in result sets of different configurations only judged once
  - All judgments by the same person

- **Measures**
  - $P(10)$, $P(20)$, $P(30)$
  - infAP [Yilmaz and Aslam, 2006]
  - calculated using trec_eval
Evaluation results

<table>
<thead>
<tr>
<th>Conf.</th>
<th>P(10)</th>
<th>P(20)</th>
<th>P(30)</th>
<th>infAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>conf_1</td>
<td>0.2418</td>
<td>0.2051</td>
<td>0.1700</td>
<td>0.1484</td>
</tr>
<tr>
<td>conf_2</td>
<td>0.3089</td>
<td>0.2778</td>
<td>0.2502</td>
<td>0.2487</td>
</tr>
<tr>
<td>conf_3</td>
<td>0.3165</td>
<td>0.2608</td>
<td>0.2131</td>
<td>0.2114</td>
</tr>
<tr>
<td>conf_4</td>
<td>0.3114</td>
<td>0.2582</td>
<td>0.2097</td>
<td>0.2001</td>
</tr>
<tr>
<td>conf_5</td>
<td>0.3114</td>
<td>0.2582</td>
<td>0.2097</td>
<td>0.2000</td>
</tr>
<tr>
<td>conf_6</td>
<td>0.3848</td>
<td>0.3405</td>
<td>0.3046</td>
<td>0.3253</td>
</tr>
<tr>
<td>conf_7</td>
<td>0.3924</td>
<td>0.3494</td>
<td>0.3089</td>
<td>0.3326</td>
</tr>
<tr>
<td>conf_8</td>
<td>0.3911</td>
<td>0.3487</td>
<td>0.3080</td>
<td>0.3318</td>
</tr>
</tbody>
</table>
## Evaluation results - query expansion

<table>
<thead>
<tr>
<th>Conf.</th>
<th>P(10)</th>
<th>P(20)</th>
<th>P(30)</th>
<th>infAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>conf_1</td>
<td>0.2418</td>
<td>0.2051</td>
<td>0.1700</td>
<td>0.1484</td>
</tr>
<tr>
<td>conf_2</td>
<td>0.3089</td>
<td>0.2778</td>
<td>0.2502</td>
<td>0.2487</td>
</tr>
<tr>
<td>conf_3</td>
<td>0.3165</td>
<td>0.2608</td>
<td>0.2131</td>
<td>0.2114</td>
</tr>
<tr>
<td>conf_4</td>
<td>0.3114</td>
<td>0.2582</td>
<td>0.2097</td>
<td>0.2001</td>
</tr>
<tr>
<td>conf_5</td>
<td>0.3114</td>
<td>0.2582</td>
<td>0.2097</td>
<td>0.2000</td>
</tr>
<tr>
<td>conf_6</td>
<td>0.3848</td>
<td>0.3405</td>
<td>0.3046</td>
<td>0.3253</td>
</tr>
<tr>
<td>conf_7</td>
<td>0.3924</td>
<td>0.3494</td>
<td>0.3089</td>
<td>0.3326</td>
</tr>
<tr>
<td>conf_8</td>
<td>0.3911</td>
<td>0.3487</td>
<td>0.3080</td>
<td>0.3318</td>
</tr>
</tbody>
</table>
### Evaluation results - result expansion

<table>
<thead>
<tr>
<th>Conf.</th>
<th>P(10)</th>
<th>P(20)</th>
<th>P(30)</th>
<th>infAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>conf_1</td>
<td>0.2418</td>
<td>0.2051</td>
<td>0.1700</td>
<td>0.1484</td>
</tr>
<tr>
<td>conf_2</td>
<td>0.3089</td>
<td>0.2778</td>
<td>0.2502</td>
<td>0.2487</td>
</tr>
<tr>
<td>conf_3</td>
<td>0.3165</td>
<td>0.2608</td>
<td>0.2131</td>
<td>0.2114</td>
</tr>
<tr>
<td>conf_4</td>
<td>0.3114</td>
<td>0.2582</td>
<td>0.2097</td>
<td>0.2001</td>
</tr>
<tr>
<td>conf_5</td>
<td>0.3114</td>
<td>0.2582</td>
<td>0.2097</td>
<td>0.2000</td>
</tr>
<tr>
<td>conf_6</td>
<td>0.3848</td>
<td>0.3405</td>
<td>0.3046</td>
<td>0.3253</td>
</tr>
<tr>
<td>conf_7</td>
<td>0.3924</td>
<td>0.3494</td>
<td>0.3089</td>
<td>0.3326</td>
</tr>
<tr>
<td>conf_8</td>
<td>0.3911</td>
<td>0.3487</td>
<td>0.3080</td>
<td>0.3318</td>
</tr>
</tbody>
</table>
Evaluation results - nearly identical

<table>
<thead>
<tr>
<th>Conf.</th>
<th>P(10)</th>
<th>P(20)</th>
<th>P(30)</th>
<th>infAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>conf_1</td>
<td>0.2418</td>
<td>0.2051</td>
<td>0.1700</td>
<td>0.1484</td>
</tr>
<tr>
<td>conf_2</td>
<td>0.3089</td>
<td>0.2778</td>
<td>0.2502</td>
<td>0.2487</td>
</tr>
<tr>
<td>conf_3</td>
<td>0.3165</td>
<td>0.2608</td>
<td>0.2131</td>
<td>0.2114</td>
</tr>
<tr>
<td>conf_4</td>
<td>0.3114</td>
<td>0.2582</td>
<td>0.2097</td>
<td>0.2001</td>
</tr>
<tr>
<td>conf_5</td>
<td>0.3114</td>
<td>0.2582</td>
<td>0.2097</td>
<td>0.2000</td>
</tr>
<tr>
<td>conf_6</td>
<td>0.3848</td>
<td>0.3405</td>
<td>0.3046</td>
<td>0.3253</td>
</tr>
<tr>
<td>conf_7</td>
<td>0.3924</td>
<td>0.3494</td>
<td>0.3089</td>
<td>0.3326</td>
</tr>
<tr>
<td>conf_8</td>
<td>0.3911</td>
<td>0.3487</td>
<td>0.3080</td>
<td>0.3318</td>
</tr>
</tbody>
</table>
Discussion of P(n)

- Buckley and Voorhees, 2000] suggest that 50 queries should be used for P(30)
- For P(n<30) number of queries should be increased
- 100 queries are suggested for P(20)

- We use 79 queries for P(10), P(20) and P(30)
- Ranking of system configurations is identical for P(20), P(30) and infAP
Discussion of infAP (1)

- TREC 8 Ad-Hoc collection
  - 528,155 documents * 50 queries = 26,407,750 possible relevance judgments
  - 86,830 query documents pairs judged
    - Depth-100 pooling of 129 systems
    - 0.33% of all possible relevance judgments performed

- Our collection
  - 1026 documents * 79 queries = 81,054 possible relevance judgments
  - 1938 query documents pair judged
    - Depth-30 pooling of 8 configurations
    - 498 additional judgments
    - 2.39% of all possible relevance judgments performed
Discussion of infAP (2)

- Depth-100 pool would be 4138 query document pairs
- We judged 46.83% of potential depth-100 pool
- [Yilmaz and Aslam, 2006] find that with 25% of judgments ranking of systems is identical to 100% in TREC 8 Ad-Hoc
- Ranking of system configurations is identical for P(20), P(30) and infAP
Thank you for your attention!

- Questions / Comments?
- pscheir@know-center.at
- http://www.aposdle.org/
Our questions

- **What do you think about**
  - the size of the collection?
  - the amount of queries?
  - the amount of judgments?
  - the measures that were used?
Associative Network

- Exact search
Associative Network

- Exact search
Associative Network

- **Exact search**
Associative Network

- Exact search

![Diagram of Associative Network]

Concept layer

Document layer
Associative Network

- **Associative search**
Associative Network

- Associative search
 Associative Network

- Associative search
Associative Network

- Associative search
Associative Network

- Associative search

Associative Information Retrieval

- Find relevant information by retrieving information that is by some means *associated* with information already known to be relevant

- **Not only more but more relevant**
Semantic Similarity

$$sim(c_1, c_2) = \frac{2 \cdot lcs(c_1, c_2)}{depth(c_1) + depth(c_2)}$$

(Wu & Palmer, 1994)

$c_1$ ... first concept
$c_2$ ... second concept
$lcs$ ... least common subsumer of two concepts
$depth$ ... depth of concept in the class hierarchy
Textual Similarity

\[ \text{sim}(d1, d2) = \text{score}(d_{125}, d2) \]

\(d1\) ... document vector of the first document
\(d2\) ... document vector of the second document
\(d_{125}\) ... document vector of the first document with all term weights removed except the 25 highest terms weights

\[ \text{score}(q, d) = \text{coord}(q, d) \cdot \text{queryNorm}(q) \]
\[ \cdot \sum_{t \in q} (t.f(t_{in}d) \cdot \text{idf}(t)^2 \cdot t.getBoost() \cdot \text{norm}(t, d)) \]

More details: Javadoc of `org.apache.lucene.search.Similarity`