Broad Expertise Retrieval in Sparse Data Environments
Krisztian Balog, Toine Bogers, Leif Azzopardi
Maarten de Rijke, Antal van den Bosch

Introduction
- Launch of the Expert Finding task at TREC generated a lot of interest in expertise retrieval
  - TREC introduced and used the W3C collection in 2005 and 2006
  - CSIRO collection in 2007
- Nearly all of the work has been evaluated on the W3C collection
- We focus on expertise retrieval in a different setting

Introduction (2)
- Our setting: intranets of universities and other knowledge-intensive organizations
  - Data used crawled from the website of Tilburg University
- Differs from the W3C (and CSIRO) setting(s)
  - Clean, structured, and focused, but limited number of documents
  - Contains information on topic and organizational hierarchy
  - Bilingual (English and Dutch)
  - List of expertise areas of an individual provided by the person him-/herself

Outline
- Introduction
- Tasks
- UvT Expert Collection
- Baseline models
- Advanced models
- Conclusions

Tasks
- Expert finding
  - Who are the experts on topic X?
- Expert profiling
  - What topics does a candidate know about?
  - I.e., quantifying the competence of a person on a list of topics (knowledge areas)

Probabilistic retrieval framework
- Expert finding
  - $p(c|a)$ — What is the probability of a candidate ca being an expert given the query topic q?
- Expert profiling
  - $p(g|a)$ — What is the probability of a knowledge area (topic) being part of the candidate’s (expertise) profile?
- Approach
  - Strength of the association between a topic and a person is indication of the level of expertise
  - $p(c|q) = \frac{p(q|ca)p(ca)}{p(q)} = p(c|q) \times \frac{p(c|a)}{p(a)}$
Example

Outline

- Introduction
- Tasks
  - UvT Expert Collection
- Baseline models
- Advanced models
- Conclusions

Document types

1. research descriptions (RD)

Document types

2. course descriptions (CD)

Document types

3. publications (PUB)

Document types

4. personal homepages (HP)
Statistics of the UvT Expert Collection

<table>
<thead>
<tr>
<th></th>
<th>Dutch</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td># experts</td>
<td>1168</td>
<td>981</td>
</tr>
<tr>
<td># topics</td>
<td>1491</td>
<td>981</td>
</tr>
<tr>
<td>Avg # of topics/expert</td>
<td>5.8</td>
<td>3.3</td>
</tr>
<tr>
<td># experts with HP</td>
<td>318</td>
<td>313</td>
</tr>
<tr>
<td># experts with CD</td>
<td>318</td>
<td>313</td>
</tr>
<tr>
<td># experts with PUB</td>
<td>734</td>
<td>55</td>
</tr>
<tr>
<td># topics</td>
<td>1168</td>
<td>1491</td>
</tr>
</tbody>
</table>

Table 2

W3C vs UvT

<table>
<thead>
<tr>
<th></th>
<th>W3C</th>
<th>UvT</th>
</tr>
</thead>
<tbody>
<tr>
<td>documents</td>
<td>~300.000</td>
<td>research-, course descri.,</td>
</tr>
<tr>
<td></td>
<td>web, lists, wiki</td>
<td>publications, homepages</td>
</tr>
<tr>
<td>languages</td>
<td>English</td>
<td>English/Dutch</td>
</tr>
<tr>
<td>candidates</td>
<td>1066</td>
<td>1168</td>
</tr>
<tr>
<td># topics</td>
<td>1491</td>
<td>981</td>
</tr>
<tr>
<td></td>
<td>evidence</td>
<td>English/Dutch</td>
</tr>
</tbody>
</table>

Research questions

- How do state-of-the-art algorithms developed on the W3C collection perform in this new setting?
- More generally: Do the lessons from the Expert Finding task at TREC carry over?
- How does the inclusion or exclusion of different document types affect expertise retrieval tasks?
- How can topical and organizational structure be used for retrieval purposes?

Outline

- Introduction
- Tasks
- UvT Expert Collection
- Baseline models
- Advanced models
- Conclusions

Approach

- Problem: how to estimate \( p(q|c) \)?
- Approach: association finding between topics and people
- Candidate-based approaches
  - Create textual model of candidates’ knowledge according to documents they are associated with
  - Model 1: query likelihood
  - Model 3: KL-divergence of candidate and topic models
- Document-based approach
  - Model 2: Find out who is most strongly associated with the documents that best describe the topic
- Estimate document-candidate associations

Candidate-based approaches

- Collect all term information from all documents associated with given candidate
- Smooth it with a background model
- Use this to represent candidate

\[
p(q|c) = (1 - \lambda) \cdot p(q|u) + \lambda \cdot p(t) \]

\[
p(t|u) = \sum_d p(t|d) \cdot p(d|u)
\]
Candidate-based approaches (2)

- Model 1 (Candidate model)
  - Query-likelihood
    How likely is that a candidate would produce the query?
    \[ p(q|\theta_c) = \prod_{t \in Q} p(t|\theta_c)^{n(t,q)} \]

- Model 3 (Topic model)
  - Topic representation using relevance models
  - KL-divergence between topic and candidate models
    \[ K_L(\theta_c || \theta) = \sum_t p(t|\theta_c) \log \frac{p(t|\theta_c)}{p(t|\theta)} \]

Document-based approach

- Find documents relevant to the query
- Find out who is most strongly associated with the relevant documents
  \[ p(q|\text{rev}) = \sum_d p(q|d)p(d|\text{rev}) \]
  \[ p(q|\theta_c) = \prod_{t \in Q} p(t|\theta_c)^{n(t,q)} \]
  \[ p(t|\theta_c) = (1 - \lambda) \cdot p(t|d) + \lambda \cdot p(t) \]

Document-candidate associations

- Need to estimate \( p(d|ca) \), the strength of the association between document \( d \) and candidate \( ca \)
- In our setting, authors of documents can unambiguously identified
  \[ p(d|ca) = 1 \text{ if } ca \text{ is author of document } d, 0 \text{ otherwise} \]

Evaluation

<table>
<thead>
<tr>
<th>Document Types</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.85</td>
<td>0.90</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>Portuguese</td>
<td>0.75</td>
<td>0.80</td>
<td>0.85</td>
<td>0.88</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.65</td>
<td>0.70</td>
<td>0.75</td>
<td>0.80</td>
</tr>
<tr>
<td>German</td>
<td>0.55</td>
<td>0.60</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>Italian</td>
<td>0.45</td>
<td>0.50</td>
<td>0.55</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Document Types</th>
<th>Expert Profiling</th>
<th>Publication</th>
<th>Course Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>90%</td>
<td>80%</td>
<td>70%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>80%</td>
<td>70%</td>
<td>60%</td>
</tr>
<tr>
<td>Spanish</td>
<td>70%</td>
<td>60%</td>
<td>50%</td>
</tr>
<tr>
<td>German</td>
<td>60%</td>
<td>50%</td>
<td>40%</td>
</tr>
<tr>
<td>Italian</td>
<td>50%</td>
<td>40%</td>
<td>30%</td>
</tr>
</tbody>
</table>

Table 1

Findings

- Models
  - Model 2 (document-based approach) performs best across the board
  - When the data is clean and very focused (RD), Model 3 (Topic model) may outperform it

- Document types
  - Expert profiling benefits much from the clean data (research and course descriptions)
  - Publications contribute the most to the expert finding task
  - Adding more data helps, however adding the homepages does not prove to be particularly useful
Findings (2)

- Compare scores across collections (Model 2, MAP)

<table>
<thead>
<tr>
<th></th>
<th>TREC 05</th>
<th>TREC 06</th>
<th>UvT 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding</td>
<td>0.22</td>
<td>0.47</td>
<td>0.28 (UK)</td>
</tr>
<tr>
<td>Profiling</td>
<td>0.44</td>
<td>0.55</td>
<td>0.28 (UK)</td>
</tr>
</tbody>
</table>

- Reason: data sparseness
  - "not enough data"
  - large number of topics
  - assessments may also be sparse

Outline

- Introduction
- Tasks
- UvT Expert Collection
- Baseline models
  - Advanced models
    - Exploiting knowledge area similarity
    - Contextual information
    - Multilingual models
- Conclusions

Exploiting knowledge area similarity

- Use similar queries (q') to support the original query (q)
  - Interpolating between p(q|ca) and the further supporting evidence from all similar requests q'

\[ p'(q|ca) = \lambda p(q|ca) + (1 - \lambda) \sum_{q'} \frac{p(q'|ca)}{p(q'|ca)} \]

- Similarity between topics q and q'

Estimating knowledge area similarity

- Content-based approach
  - Kullback-Leibler divergence (KL)
  - Pointwise Mutual Information (PMI)
  - Log-likelihood (LL)
- Structure-based approach
  - Distance within the topic hierarchy (HDIST)
- Similarity scores are normalized to form probabilities

Research questions

- Does exploiting the knowledge area similarity improve effectiveness?
- Which of the various methods for capturing word relationships is most effective?

Results

- Improvement over the baseline for both tasks and languages

<table>
<thead>
<tr>
<th></th>
<th>Improvement (MAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding</td>
<td>+3.7% ... +7.7%</td>
</tr>
<tr>
<td>Profiling</td>
<td>+7.0% ... +18.6%</td>
</tr>
</tbody>
</table>

- LL method performed best
- Content-based approaches performed consistently better than HDIST
- Upshot: Method can be applied where topic hierarchy is not available
**Contextual information**

- Two-level hierarchy of organizational units (faculties and institutes) is available
- Organizational units are regarded as a context so as to compensate for data sparseness

\[ p(q, ou) = \left(1 - \sum_{ou \in OU} \lambda_{ou} \right) \cdot p(q) + \sum_{ou \in OU} \lambda_{ou} \cdot p(q|ou) \]

- \( \text{OU(ou): set of organizational units, of which candidate ca is member of} \)

**Evaluation**

- **RQ: Is our way of bringing in contextual information useful?**
- **Evaluation is done in two steps**
  - **Step 1:** Evaluation of the models of organizational units
  - **Step 2:** Combination of the contextual models with the candidate models

**Evaluation (Step 1)**

- **Evaluation of the models of organizational units**

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>NDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Finding</strong></td>
<td>0.42</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>Profiling</strong></td>
<td>0.16</td>
<td>0.32</td>
</tr>
</tbody>
</table>

- Organizational unit is relevant for a given topic (and vice versa) if at least one member of the unit selected the given topic as an expertise area

**Evaluation (Step 2)**

- **Combination of the contextual models with the candidate models**

<table>
<thead>
<tr>
<th></th>
<th>Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Finding</strong></td>
<td>-1.5% to +11.4%</td>
</tr>
<tr>
<td><strong>Profiling</strong></td>
<td>+1.9% to +25.4%</td>
</tr>
</tbody>
</table>

- **Findings**
  - Little or no improvement for Model 2
  - Clearly beneficial for Model 1 and 3

**Multilingual model**

- For knowledge institutes in Europe a multilingual (or at least bilingual) setting is typical
- Simple multilingual model
  - No spill-over of expertise/profiles across language boundaries

\[ p'(q|\alpha) = \sum_{l \in L} \lambda_l p(q|\alpha_l) \]

- \( \text{L: set of languages, translation of q to language l} \)
Evaluation

- RQ: Is a simple way of combining monolingual scores sufficient for obtaining significant improvements?
- Two languages (EN, NL), with equal weights
- Improvement over all measures for both tasks

<table>
<thead>
<tr>
<th>Improvement (MAP)</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>+14% ... +34%</td>
<td>Rating</td>
</tr>
<tr>
<td>+11% ... +79%</td>
<td>Profiling</td>
</tr>
</tbody>
</table>

Conclusions

- Summary
  - Expertise retrieval in a new setting of a typical knowledge-intensive organization
  - Introduced (and released) the UvT Expert Collection
  - Evaluated how state-of-the-art models for expertise finding performed in this new setting
  - Refined models in order to exploit the different characteristics within the data environment (data, topicality, org. structure)
- Findings
  - Current models generalize well
  - Refined models result in significant improvement

Next steps

- Ask people to manually assess the profiles the system generated for them
- Use it as a ‘recommendation system’ (save time on filling out web forms)
- Use the data/models for other institutes where these types of facilities are not available

Questions?

www.science.uva.nl/~kbalog

UvT Expert collection
http://ilk.uvt.nl/uvt-expert-collection/