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 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

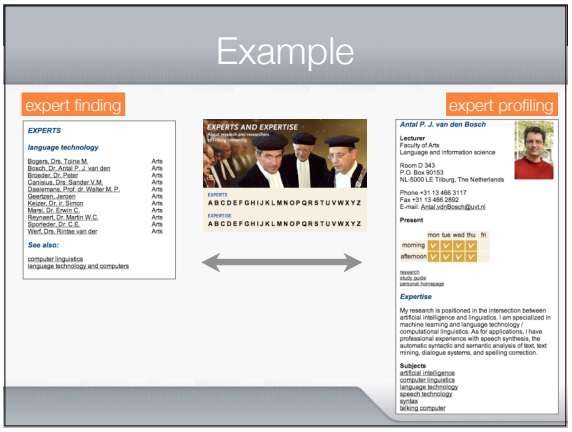
- Expert finding
  - *Who are the experts on topic X?*
- Expert profiling
  - *What topics does a candidate know about?*
  - I.e., quantifying the competence of a person on a list of topics (knowledge areas)

Probabilistic retrieval framework

- Expert finding
  - ***p(ca|q)*** — *What is the probability of a candidate ca being an expert given the query topic q?*
- Expert profiling
  - ***p(q|ca)*** — *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

$$p(ca|q) = \frac{p(q|ca)p(ca)}{p(q)} \Rightarrow p(ca|q) \propto \boxed{p(q|ca)}p(ca)$$

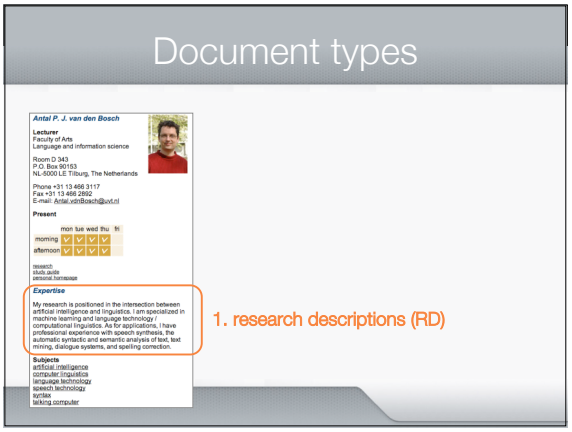
Example



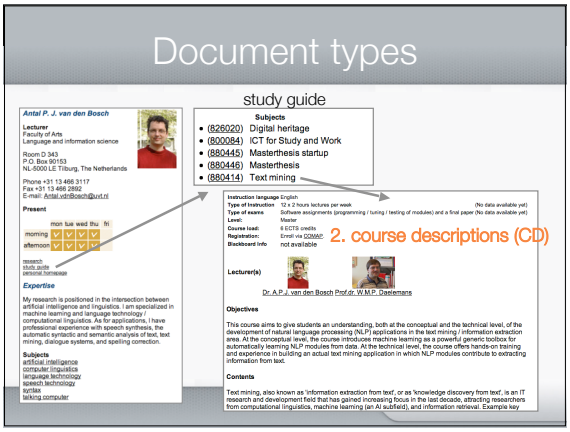
Outline

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

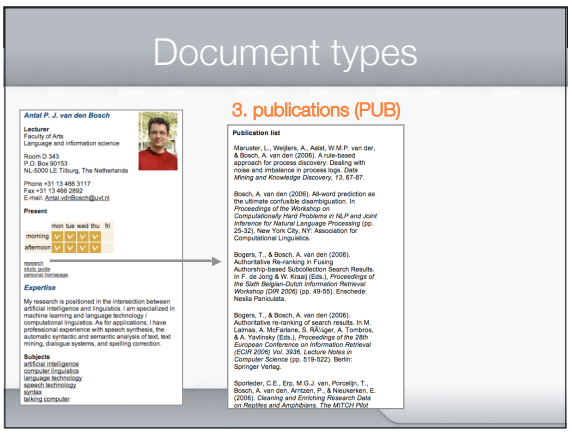
Document types



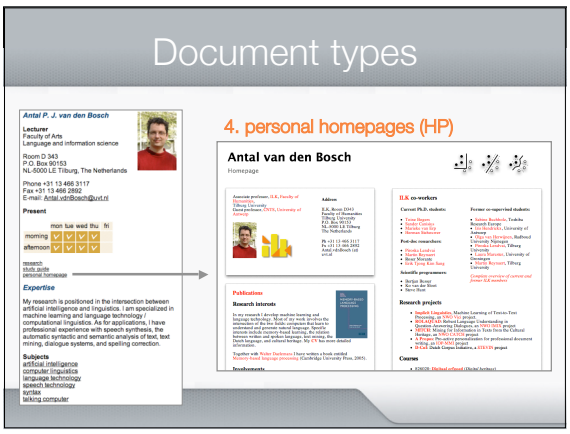
Document types



Document types



Document types



Statistics of the UvT Expert Collection

	Dutch	English
# experts	1168	
# topics	1491	981
Avg. no. of topics/expert	5.8	5.9
Avg. no of experts/topic	2.9	3.3
# experts with RD	329	313
# experts with CD	318	
# experts with PUB	734	
# experts with HP	318	

Table 2

W3C vs UvT

	W3C	UvT
documents	www, lists, wikis ~330.000	research-, course descr., publications, homepages ~38.000
languages	English	English/Dutch
candidates	1092	1168
# topics	(2005) 50; (2006) 55	(NL) 1491; (UK) 981
assessment	(2005) heuristic [wg] (2006) manual [evidence]	self
organizational structure		

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 $p(q|ca)$?
- 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

$$p(t|\theta_{ca}) = (1-\lambda)\cdot p(t|ca) + \lambda \cdot p(t)$$

$$p(t|ca) = \sum_d p(t|d)\cdot p(d|ca)$$

Candidate-based approaches (2)

- Model 1 (Candidate model)
  - Query-likelihood
    How likely is that a candidate would produce the query?

$$p(q|\theta_{ca}) = \prod_{t \in q} p(t|\theta_{ca})^{n(t,q)}$$

- Model 3 (Topic model)
  - Topic representation using relevance models
  - KL-divergence between topic and candidate models

$$KL(\theta_k || \theta_{ca}) = \sum_t p(t|\theta_k) \log \frac{p(t|\theta_k)}{p(t|\theta_{ca})}$$

Overview

- Baseline models
  - **Candidate-based** approaches
  - **Document-based** approach
  - Estimate document-candidate associations
  - Evaluation

Document-based approach

- Find documents relevant to the query
- Find out who is most strongly associated with the relevant documents

$$p(q|ca) = \sum_d p(q|d)p(d|ca)$$

$$\downarrow$$

$$p(q|\theta_d) = \prod_{t \in q} p(t|\theta_d)^{n(t,q)}$$

$$\downarrow$$

$$p(t|\theta_d) = (1-\lambda) \cdot p(t|d) + \lambda \cdot p(t)$$

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 if *ca* is author of document *d*, 0 otherwise

Evaluation

	Expert finding						Expert profiling					
Document types	Model 1			Model 2			Model 3			Model 1			Model 2			Model 3		
	%q	MAP	MRR	%q	MAP	MRR	%q	MAP	MRR	%ca	MAP	MRR	%ca	MAP	MRR	%ca	MAP	MRR
English																		
RD	**97.8**	0.126	0.269	83.5	**0.144**	**0.311**	83.3	0.129	0.271	**100**	0.089	0.189	39.3	**0.232**	**0.465**	41.1	0.166	0.337
CD	**97.8**	0.118	0.227	91.7	**0.123**	**0.248**	91.7	0.118	0.226	**32.8**	0.188	0.381	32.4	0.195	**0.385**	32.7	**0.203**	0.370
PUB	97.8	0.200	0.330	**98.0**	**0.216**	**0.372**	**98.0**	0.145	0.257	**78.9**	0.167	0.364	74.5	**0.212**	**0.442**	**78.9**	0.135	0.299
HP	**97.8**	**0.081**	**0.186**	97.4	0.071	0.168	97.2	0.062	0.149	**31.2**	0.150	0.299	28.8	**0.185**	**0.335**	30.1	0.136	0.287
RD+CD	**97.8**	0.188	0.352	92.9	**0.193**	**0.360**	92.9	0.150	0.273	**100**	0.145	0.286	61.3	**0.251**	**0.477**	63.2	0.217	0.416
RD+CD+PUB	97.8	0.235	0.373	**98.1**	**0.277**	**0.439**	**98.1**	0.178	0.305	**100**	0.196	0.380	87.2	**0.280**	**0.533**	89.5	0.170	0.344
RD+CD+PUB+HP	97.8	0.237	0.372	**98.6**	**0.280**	**0.441**	98.5	0.166	0.293	**100**	0.199	0.387	88.7	**0.281**	**0.525**	90.9	0.169	0.329
Dutch																		
RD	**61.3**	0.094	0.229	38.4	**0.137**	**0.336**	38.3	0.127	0.295	**38.0**	0.127	0.386	34.1	**0.138**	**0.420**	**38.0**	0.105	0.327
CD	**61.3**	0.107	0.212	49.7	0.128	0.256	49.7	**0.136**	**0.261**	**32.5**	0.151	0.389	31.8	0.158	**0.396**	**32.5**	**0.170**	0.380
PUB	**61.3**	0.193	0.319	59.5	**0.218**	**0.368**	59.4	0.173	0.291	**78.8**	0.126	0.364	76.0	**0.150**	**0.424**	**78.8**	0.103	0.294
HP	**61.3**	0.063	0.169	56.6	**0.064**	**0.175**	56.4	0.062	0.163	**29.8**	0.108	0.308	27.8	**0.125**	**0.338**	**29.8**	0.098	0.255
RD+CD	**61.3**	0.159	0.314	51.9	**0.184**	**0.360**	51.9	0.169	0.324	**60.5**	0.151	0.410	57.2	**0.166**	**0.431**	60.4	0.159	0.384
RD+CD+PUB	61.3	0.244	0.398	**61.5**	**0.260**	**0.424**	61.4	0.210	0.350	**90.3**	0.165	0.445	88.2	**0.189**	**0.479**	**90.3**	0.126	0.339
RD+CD+PUB+HP	61.3	0.249	0.401	**62.6**	**0.265**	**0.436**	**62.6**	0.195	0.344	**91.9**	0.164	0.426	90.1	**0.195**	**0.488**	**91.9**	0.125	0.328

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)

	TREC'05	TREC'06	UvT
Finding	0.22	0.47	0.28 (UK) 0.26 (NL)
Profiling	0.44	0.55	0.28 (UK) 0.19 (NL)

- 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'*

$$p'(q|ca) = \lambda p(q|ca) + (1-\lambda) \sum_{q'} \boxed{p(q|q')} p(q'|ca)$$

similarity between topics *q* and *q'*

Estimating knowledge area similarity

- Content-based approach
  - Kullback-Leiber 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

	Improvement (MAP)
Finding	+3.7% … +7.7%
Profiling	+7.0% … +18.6%

- 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

$$p'(q|ca) = (1 - \sum_{ou \in OU(ca)} \lambda_{ou}) \cdot p(q|ca) + \sum_{ou \in OU(ca)} \lambda_{ou} \cdot p(q|ou)$$

?

OU(ca): set of organizational units, of which candidate *ca* is member of

Contextual information (2)

- *p(q|ou)*: strength of the association between query *q* and organizational unit *ou*
- Estimated using baseline models
  - Substitute *ca* with *ou*
  - Document-organizational unit associations *p(d|ou)*
    - An organizational unit is associated with all the documents that its members have authored

$$p(d|ou) = max_{ca \in ou} p(d|ca)$$

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

	MAP	MRR
Finding	0.42-0.70	0.54-0.83
Profiling	0.18-0.52	0.55-0.77

- 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

	Improvement (MAP)
Finding	-1.5% … +11.4%
Profiling	+1.8% … +25.4%

- 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

language-specific smoothing parameter $\sum_{l \in L} \lambda_l = 1$

$$p'(q|ca) = \sum_{l \in L} \lambda_l p(q_l|ca)$$

translation of q to language l

set of languages

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

	Improvement (MAP)
Finding	+14% … +34%
Profiling	+11% … +79%

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/