

Formal Models for Expert Finding in Enterprise Corpora

Krisztian Balog, Leif Azzopardi*, and Maarten de Rijke
ISLA, University of Amsterdam
* University of Strathclyde, Glasgow

1


Motivation

- Searching an organization's document repositories
- Getting to an organization's knowledge
 - » Managing the expertise of employees
 - Identifying experts in a certain area
 - People you'd contact with questions on the topic
 - » Search for people as a social tool
- From retrieving documents to retrieving objects
 - » answers, movies, books, ads, people, soccer match results, ...

2

The picture

Expert finding topic: authoring tools


Dave Pawson candidate-0319
 E-mail: dave.pawson@gmail.com, dave.pawson@virgin.net
 Homepage: http://www.dpawson.co.uk/
 Keywords: priority, authoring, tool, accessible, checkpoints, autools, guideline, checkpoint, alerts, webcontent, prompts, markup
 Profile:

authoring tool guidelines	<input type="checkbox"/>	TOP 20
web content accessibility	<input type="checkbox"/>	TOP 20
xsl extensible stylesheet lang...	<input type="checkbox"/>	
mobile web initiative workshop...	<input type="checkbox"/>	
wcag reviewers	<input type="checkbox"/>	
more...	<input type="checkbox"/>	

Find more about this person on: [Google](#) | [CiteSeer](#) | [Portal.acm.org](#)

Given a topic, find experts on the topic

3

Expert finding

- TREC 2005 Enterprise Track
 - » Common platform to empirically assess methods and techniques
- Why an interesting task?
 - » Focused retrieval
 - » Structure
 - Collection, links, documents, ...
 - » Bring in more and more "non-factual" aspects of relevancy
 - Importance, expertise, ...

4

Overview

- Two models implementing different intuitions about expert finding
- Document-candidate associations
- Research questions addressed so far
- Results
- Conclusions

5

Modeling Expert Search

- Given a crawl of an intranet, a list of candidate experts, and a set of topics, find the experts for each of the topics
- Rank candidates according to the probability of a candidate being an expert
 - » **Model 1:** Create a textual model of candidates' knowledge according to the documents with which they are associated
 - » **Model 2:** Find out who is most strongly associated with the documents that best describe the topic

6

Approach

- "What is the probability of a candidate ca being an expert given the topic q ?"
- Use Bayes to rewrite this

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

- Ranking of candidates proportional to the probability of a query given a candidate
- Model 1: compute $p(q|ca)$ by building a candidate model from docs associated with candidate, and generate the query from this model
- Model 2: q and ca conditionally independent, resolve relation through document-candidate associations

7

2 types of Doc-Cand associations

- Need: estimate the probability that a doc is associated with a candidate
- Assume: extraction component produces associations $a(d,ca)$ for each d and ca

1. Document-centric association:



$$p(d|ca) = \frac{a(d,ca)}{\sum_{c \in C} a(d,c)}$$

- "Use this to rank candidates for a fixed d , to find candidate that made the biggest contribution to d "

8

Doc-Cand associations (2)

- Assume: extraction component produces associations $a(d,ca)$ for each d and ca

2. Candidate-centric association:



$$p(ca|d) = \frac{a(d,ca)}{\sum_{dx \in D} a(dx,ca)}$$

- "Use this to rank docs for a fixed ca , to find doc most strongly associated with ca "

9

Computing $a(d,ca)$

- Restricted information extraction task
- Given doc d and candidate ca (ID, name(s), email(s))
- Simple rule-based approach
 - A0: Exact match: 1, if candidate name occurs as specified
 - A1: Name match: 1, if last name plus initial of the first name occurs
 - A2: Last name match: 1, if last name occurs
 - A3: Email match: 1, if email address occurs

$$a(d,ca) = \sum_{i=0}^k \pi_i A_i(d,ca)$$

10

Model 1

- Idea: collect all term information from all documents associated with given candidate; use this to represent candidate
- "How likely is that a candidate would produce the query?"

$$p(q|\theta_{ca}) = \prod_{t \in q} \left\{ (1 - \lambda) \left(\sum_d p(t|d) f(d,ca) \right) + \lambda p(t) \right\}^{n(t,q)}$$

representation of a candidate
background model
↓
 $p(d|ca)$ or $p(ca|d)$

11

Model 2

- Intuition: Find out who is most strongly associated with the documents that best describe the topic
- Given collection ranked according to query, examine each doc, if relevant, see who is associated with it

$$p(q|ca) = \sum_d \left\{ \prod_{t \in q} \left((1 - \lambda) p(t|d) + \lambda p(t) \right)^{n(t,q)} \right\} f(d,ca)$$

relevance of a document
 $p(d|ca)$ or $p(ca|d)$

12

Research questions

- How do Model 1 and Model 2 compare?
- What is a more effective way of capturing the strength of an association between a document and a candidate?
 - Document-centric vs candidate-centric
- ... more questions are asked (and answered) in the paper

13

Experimental set-up

- TREC Enterprise 2005 platform
- W3C collection
 - Mixture of document types crawled from w3c.org (www, wikis, e-mail lists archive, etc.)
 - All handled and processed as HTML documents
 - 330.000 documents, 5.7 GB
- List of 1092 candidate experts
 - unique ID, name, e-mail address(es)
- 50 topics, and relevance judgments
- Evaluation measures: MAP, P@10, P@20, and MRR

14

Model 1 vs Model 2

#rel	MAP	R-prec	MRR	P10	P20	
Model 1 (candidate model):						
document-centric:						
A0	492	0.1221	0.1576	0.3574	0.236	0.209
A1	486	0.1138	0.1537	0.3246	0.214	0.204
A2	447	0.0919	0.1476	0.3288	0.206	0.175
A3	424	0.1234	0.1778	0.4096	0.262	0.194
candidate-centric:						
A0	511	0.1253	0.1914	0.2759	0.236	0.227
A1	507	0.1189	0.1851	0.2537	0.216	0.206
A2	471	0.0951	0.1654	0.2604	0.202	0.186
A3	430	0.1347	0.1819	0.4839	0.280	0.208
Model 2 (document model):						
document-centric:						
A0	560	0.1731	0.2245	0.4783	0.284	0.241
A1	554	0.1670	0.2243	0.4590	0.280	0.237
A2	513	0.1294	0.1902	0.4195	0.228	0.214
A3	430	0.1222	0.1768	0.4770	0.238	0.187
candidate-centric:						
A0	580	0.1880	0.2332	0.5149	0.316	0.260
A1	575	0.1790	0.2262	0.4958	0.296	0.254
A2	543	0.1537	0.2173	0.4872	0.274	0.235
A3	439	0.1337	0.1934	0.4898	0.256	0.205

Table 2

15

Model 1 vs Model 2

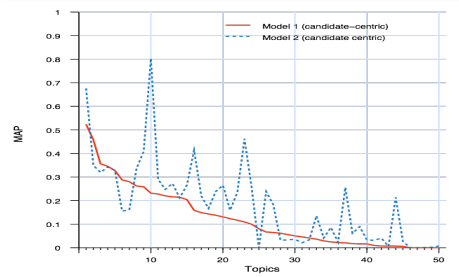


Figure 2

16

Centricity

#rel	MAP	R-prec	MRR	P10	P20	
Model 1 (candidate model):						
document-centric:						
A0	492	0.1221	0.1576	0.3574	0.236	0.209
A1	486	0.1138	0.1537	0.3246	0.214	0.204
A2	447	0.0919	0.1476	0.3288	0.206	0.175
A3	424	0.1234	0.1778	0.4096	0.262	0.194
candidate-centric:						
A0	511	0.1253	0.1914	0.2759	0.236	0.227
A1	507	0.1189	0.1851	0.2537	0.216	0.206
A2	471	0.0951	0.1654	0.2604	0.202	0.186
A3	430	0.1347	0.1819	0.4839	0.280	0.208
Model 2 (document model):						
document-centric:						
A0	560	0.1731	0.2245	0.4783	0.284	0.241
A1	554	0.1670	0.2243	0.4590	0.280	0.237
A2	513	0.1294	0.1902	0.4195	0.228	0.214
A3	430	0.1222	0.1768	0.4770	0.238	0.187
candidate-centric:						
A0	580	0.1880	0.2332	0.5149	0.316	0.260
A1	575	0.1790	0.2262	0.4958	0.296	0.254
A2	543	0.1537	0.2173	0.4872	0.274	0.235
A3	439	0.1337	0.1934	0.4898	0.256	0.205

Table 2

17

Expert finding findings

- Recall our aim was to develop intuitive models for expert finding
 - Model 1 vs. Model 2
- Findings
 - Model 2 outperforms Model 1
 - Candidate-centric estimation generally performs better
- Other findings (see paper for details)
 - Smoothing
 - Not a single lesson for both models
 - Extraction methods, combination of extraction methods
 - The quality not the quantity of the associations is essential
 - Topically focused subset of documents
 - Improves responsiveness without hurting performance too much

18

Universiteit van Amsterdam, ISLA

Looking forward

★★★★★ **Dave Pawson** candidate-0319

E-mail: dave.pawson@gmail.com, dave.pawson@virgin.net
 Homepage: http://www.dpawson.co.uk/
 Keywords: priority, authoring, tool, accessible, checkpoints, autools, guideline, checkpoint, alerts, webcontent, prompts, markup

Profile:

authoring tool guidelines		TOP 20
web content accessibility		TOP 20
xsl extensible stylesheet lang...		
mobile web initiative workshop...		
wcag reviewers		
more...		

Find more about this person on: [Google](#) | [CiteSeer](#) | [Portal.acm.org](#)

Topical profiling: expertise areas for a given person

19

Universiteit van Amsterdam, ISLA

Profiling Experts

- Assume: a list of knowledge areas
- Need: to measure the person's competence in each of the knowledge areas ($score(ca,ka)$)
- Baseline: "invert expert finding"
 - Probability: $p_{ES}(ca|ka)$
 - Rank: $1/rank_{ES}(ca)$
- Method 1
 - Given knowledge area, sum relevancy scores (based on LMs) over all docs associated with a given candidate
- Method 2
 - Compare LMs of the knowledge area and the candidate, and measure their divergence

20

Universiteit van Amsterdam, ISLA

Profiling experts (2)

- Profiling Method 1 beats profiling Method 2 beats probabilistic baseline beats rank baseline
 - Measures used: MAP, MRR (max around .5)
- Further improvements by taking into account whether a candidate is among the top ranked experts in a given area
 - MRR maxes out around .7
- Use profiling to improve expert finding
 - Rerank ES results using profiling rank of a candidate for a knowledge area (MRR scores up by ~25%)

21

Universiteit van Amsterdam, ISLA

Questions?

Krisztian Balog
 kbalog@science.uva.nl

22

Universiteit van Amsterdam, ISLA

Extraction

method	%cand	%rel_cand	#avg	%docs
Extraction by name:				
A0: EXACT MATCH	63.74%	62.34%	466	41.27%
A1: NAME MATCH	69.32%	68.01%	468	42.23%
A2: LAST NAME MATCH	84.62%	83.96%	1023	64.17%
Extraction by e-mail address:				
A3: EMAIL MATCH	41.76%	40.06%	162	17.93%
Combining methods:				
A0 and A3:	66.03%	64.55%	552	42.59%
A1 and A3:	70.51%	69.26%	556	43.44%
A2 and A3:	85.35%	84.73%	1094	64.68%

Table 1

- EMAIL_MATCH adds little on top of name-based extraction methods

23

Universiteit van Amsterdam, ISLA

Extraction (2)

Model 2 (document model):						
document-centric:						
A0	560	0.1731	0.2245	0.4783	0.284	0.241
A1	554	0.1670	0.2243	0.4590	0.280	0.237
A2	513	0.1294	0.1902	0.4195	0.228	0.214
A3	430	0.1222	0.1768	0.4770	0.238	0.187
candidate-centric:						
A0	580	0.1880	0.2332	0.5149	0.316	0.260
A1	575	0.1790	0.2262	0.4958	0.296	0.254
A2	543	0.1537	0.2173	0.4872	0.274	0.235
A3	439	0.1337	0.1934	0.4898	0.256	0.205

Table 2

- The quality of the extraction—and not the number—of associations has the main impact on overall expert finding performance

24

Extraction (3)

	#rel	MAP	R-prec	MRR	P10	P20
Model 1 (candidate model):						
A3	430	0.1347	0.1819	0.4839	0.280	0.208
A0	511	0.1253	0.1914	0.2759	0.236	0.227
COMB	514	0.1163	0.1785	0.3358	0.190	0.199
Model 2 (document model):						
A3	439	0.1337	0.1934	0.4898	0.256	0.205
A0	580	0.1880	0.2332	0.5149	0.316	0.260
COMB	590	0.1894	0.2434	0.5043	0.316	0.260

Table 3

- Combinations of extraction methods did not improve performance on all measures

25

Model 1 vs Model 2

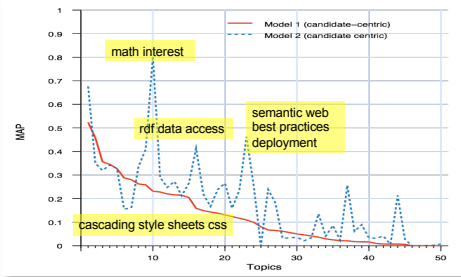


Figure 2

26