Associating People and Documents

Krisztian Balog and Maarten de Rijke

ISLA, University of Amsterdam
http://ilps.science.uva.nl

Motivation

- *Expert finding* = identifying a list of people who are knowledgeable about a given topic
  - *Who are the experts on topic X?*
- Expert finding has generated a lot of interest since the launch of the TREC Enterprise Track in 2005
- Two main families of models emerged

Motivation (2)

- Candidate models
  - Create a textual representation of candidates according to the documents with which they are *associated*
- Document models
  - Find out who is most strongly *associated* with the documents that best describe the topic
- Feature shared by many of the models: *associations* between people (candidates) and documents
- Such associations have received relatively little attention so far

Research Questions

- What is the impact of document-candidate associations on the end-to-end performance of expert finding models?
- What are effective ways of capturing the strength of the associations?
- How sensitive are expert finding models to different document-candidate association methods?

Outline

- Two models for expert finding
- Experimental setup
- Establishing associations
- Conclusions

Two Models for Expert Finding

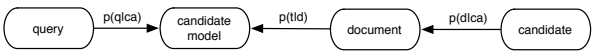
- Two principal expert finding strategies [1]
  - Candidate models
    ("profile based" or "query independent" approaches)
  - Document models
    ("query dependent" approaches)
- What is the probability of a candidate ca being an expert given the query topic q?

$$p(ca|q) \propto p(q|ca) \cdot p(ca)$$

[1] K. Balog, L. Azzopardi, M. de Rijke. **Formal models for expert finding in enterprise corpora**. In *SIGIR 2006*, pages 43-50, 2006.

Model 1: Candidate Model

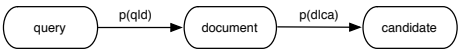
- Collect all term information from documents associated with the candidate
- Use it to represent the candidate
- *How likely is that a candidate would produce the query?*



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

Model 2: Document Model

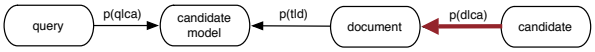
- Find documents relevant to the query
- Examine who is associated with each document



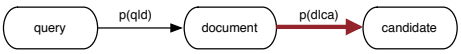
$$p(q|ca) = \sum_{d} \left\{ \prod_{t \in q} \left((1-\lambda) \cdot p(t|d) + \lambda \cdot p(t) \right)^{n(t,q)} \right\} \cdot p(d|ca)$$

Document-candidate Associations

Candidate model



Document model



Document-candidate Associations

$$p(d|ca) = \frac{p(ca|d) \cdot p(d)}{p(ca)}$$

- Reading of *p(ca|d)* is different for the two models
  - Model 1: the degree to which *ca*'s expertise is described by *d*
  - Model 2: a ranking of candidates associated with *d* based on their contribution made to *d*

Outline

- Two models for expert finding
- Experimental setup
- Establishing associations
- Conclusions

Experimental Setup

- TREC Enterprise platform
  - W3C collection
  - 2005 (50) and 2006 (49) topics
  - List of 1092 expert candidates is given
- All documents as plain text, no stemming
- Mean Average Precision

Person Name Identification

- Six match types [2]
- Candidate occurrences are replaced with a unique identifier

Type	Pattern	Example	Ambiguity (%)
MT1	Full name	Ritu Raj Tiwari Tiwari, Ritu Raj	0.0
MT2	Email name	rtiwari@nuance.com	0.0
MT3	Combined name	Tiwari, Ritu R R R Tiwari	39.92
MT4	Abbreviated name	Ritu Raj Ritu	48.90
MT5	Short name	RRT	63.96
MT6	Alias, New Mail	Ritiwari rtiwari@hotmail.com	0.46

STRICT Match Types

[2] S. Bao, H. Duan, Q. Zhou, M. Xiong, Y. Cao, and Y. Yu. **Research on Expert Search at Enterprise Track of TREC 2006**. In *TREC 2006*, 2007.

Outline

- Two models for expert finding
- Experimental setup
- Establishing associations
- Conclusions

Boolean Model of Association

- Simplest possible choice
- Associations are binary decisions
  - They exist if the candidate occurrs in the document
  - Irrespective of the number of times the person or other candidates are mentioned

$$p(ca|d) = \begin{cases} 1, & n(ca,d) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Boolean Model of Association (2)

- Two potentially unrealistic assumptions

~~1. Candidate independence~~ **Too strong?!**

  - Candidates in the document are independent of each other, all equally important

2. Position independence
  - The positions of candidates within the document are ignored

Modeling Candidate Frequencies

- **Goal**: *p(ca|d)* indicates the strength of the association between *ca* and *d*
- **Approach**:
  - Treat candidate identifiers as terms
  - How important is a candidate (term) for a given document?
  - Use term weighting schemes

Importance of a Candidate within a Document

- TF
- IDF
- TF.IDF
- Language Modeling

Experimental Results

Table 1. Candidate mentions are treated as any other term in the document. For each year-model combination the best scores are in boldface.

Method	ALL MatchTypes				STRICT MatchTypes			
	TREC 2005		TREC 2006		TREC 2005		TREC 2006	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Boolean	**.1742**	.2172	.2809	**.4511**	**.1858**	.2196	**.3075**	**.4704**
TF	.0684(3)	.2014(3)	.1726(3)	.4408	.0640(3)	.2038(2)	.1601(3)	.4485(1)
IDF	.1676	**.2480**(3)	.2488(3)	.4488	.1845	**.2512**(3)	.2736(3)	.4670
TFIDF	.1408(1)	.2227	**.2913**	.4465	.1374(2)	.2266	.2828	.4514
LM	.0676(3)	.2013(3)	.1619(3)	.4397	.0642(3)	.2031(2)	.1586(3)	.4470(1)

- ALL vs STRICT MatchTypes

Experimental Results

Table 1. Candidate mentions are treated as any other term in the document. For each year-model combination the best scores are in boldface.

Method	ALL MatchTypes				STRICT MatchTypes			
	TREC 2005		TREC 2006		TREC 2005		TREC 2006	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Boolean	**.1742**	.2172	.2809	**.4511**	**.1858**	.2196	**.3075**	**.4704**
TF	.0684(3)	.2014(3)	.1726(3)	.4408	.0640(3)	.2038(2)	.1601(3)	.4485(1)
IDF	.1676	**.2480**(3)	.2488(3)	.4488	.1845	**.2512**(3)	.2736(3)	.4670
TFIDF	.1408(1)	.2227	**.2913**	.4465	.1374(2)	.2266	.2828	.4514
LM	.0676(3)	.2013(3)	.1619(3)	.4397	.0642(3)	.2031(2)	.1586(3)	.4470(1)

- Boolean vs frequency-based approaches

Experimental Results

Table 1. Candidate mentions are treated as any other term in the document. For each year-model combination the best scores are in boldface.

Method	ALL MatchTypes				STRICT MatchTypes			
	TREC 2005		TREC 2006		TREC 2005		TREC 2006	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Boolean	**.1742**	.2172	.2809	**.4511**	**.1858**	.2196	**.3075**	**.4704**
TF	.0684(3)	.2014(3)	.1726(3)	.4408	.0640(3)	.2038(2)	.1601(3)	.4485(1)
IDF	.1676	**.2480**(3)	.2488(3)	.4488	.1845	**.2512**(3)	.2736(3)	.4670
TFIDF	.1408(1)	.2227	**.2913**	.4465	.1374(2)	.2266	.2828	.4514
LM	.0676(3)	.2013(3)	.1619(3)	.4397	.0642(3)	.2031(2)	.1586(3)	.4470(1)

- Model 1 vs Model 2

Experimental Results

Table 1. Candidate mentions are treated as any other term in the document. For each year-model combination the best scores are in boldface.

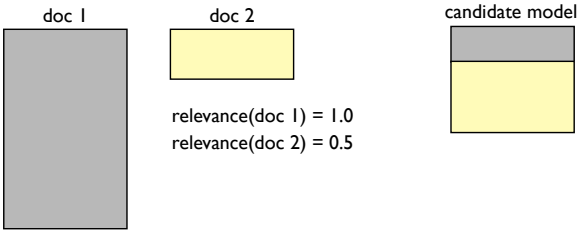
Method	ALL MatchTypes				STRICT MatchTypes			
	TREC 2005		TREC 2006		TREC 2005		TREC 2006	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Boolean	**.1742**	.2172	.2809	**.4511**	**.1858**	.2196	**.3075**	**.4704**
TF	.0684(3)	.2014(3)	.1726(3)	.4408	.0640(3)	.2038(2)	.1601(3)	.4485(1)
IDF	.1676	**.2480**(3)	.2488(3)	.4488	.1845	**.2512**(3)	.2736(3)	.4670
TFIDF	.1408(1)	.2227	**.2913**	.4465	.1374(2)	.2266	.2828	.4514
LM	.0676(3)	.2013(3)	.1619(3)	.4397	.0642(3)	.2031(2)	.1586(3)	.4470(1)

- TREC 2005, Model 2, IDF

Findings

- More beneficial to use rigid patterns for name matching (STRICT)
- Boolean method delivers excellent performance — in most cases outperforms frequency-based weighting schemes
- Model 2 is less sensitive to document-candidate associations than Model 1

Analysis



- Shorter documents contribute more to a candidate's profile

➡ need for length normalization

Using Lean Documents

- Documents contain only candidate identifiers, all other terms are filtered out
- Same weighting schemes than before

Experimental Results

Table 2. Lean document representation. For each year-model combination the best scores are in boldface.

Method	TREC 2005		TREC 2006	
	Model 1	Model 2	Model 1	Model 2
Boolean	.1858	.2196	.3075	.4704
TF	.2141(3) (+234%)	.1934 (-5.1%)	.3724(3) (+132%)	.4654 (+3.7%)
IDF	.1845	**.2512**	.2736	.4670
TFIDF	**.2304**(3) (+67.6%)	.2176 (-3.9%)	.3380(2) (+19.5%)	**.4728** (+4.7%)
LM	.2102(3) (+227%)	.1932 (-4.8%)	**.3763**(3) (+137%)	.4627 (+3.5%)

Findings

- Model 1:
  - Length normalization is needed
  - Frequency-based weighting schemes (using lean doc. representation) are preferred over the boolean model
- Model 2:
  - Length normalization is less important
  - No significant improvement over the boolean method

What do these frequency-based associations actually achieve?

Semantic Relatedness

- So far: *n(ca,d)* is the indication of the importance of a candidate given a document
- Here: alternative way of measuring the candidate's weight in the document
  - Distance between the candidate's and the document's language models

$$n'(ca,d) = \begin{cases} \text{KLDIV}(\theta_{ca}||\theta_d), & n(ca,d) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Results

Table 3. Comparing frequency-based associations using lean representations (FREQ) and semantic-relatedness of documents and candidates (SEM).

Method	TREC 2005						TREC 2006					
	Model 1			Model 2			Model 1			Model 2		
	FREQ	SEM	τ	FREQ	SEM	τ	FREQ	SEM	τ	FREQ	SEM	τ
TF	.2141	.2128	.750	.1934	.2012	.816	.3724	.3585	.761	.4654	.4590	.841
IDF	.1845	.1836	.982	.2512	.2541	.964	.2736	.2732	.986	.4670	.4586	.971
TFIDF	.2304	.2335	.748	.2176	.2269	.809	.3380	.3352	.771	.4728	.4602	.827
LM	.2102	.2117	.756	.1932	.2009	.816	.3763	.3671	.761	.4627	.4576	.841

- The correlation between FREQ and SEM is very high

➡ frequency-based associations (based on lean document representation) are capable of capturing the semantics of the associations

Outline

- Two models for expert finding
- Experimental setup
- Establishing associations
- Conclusions

Wrap-up

- Forming document-candidate associations is a key ingredient of expert finding models
- Introduced and compared a number of methods for building such associations
- Made explicit and analyzed underlying assumptions
  - Independency of candidates
  - Frequency is an indication of strength

Wrap-up (2)

- Gained insights into the inner-workings of two principal expert finding strategies
  - Candidate-based models
    - Sensitive to associations
    - Standard document representation suffers from length normalization
  - Document-based models
    - Less dependent on associations
    - Very moderate improvements over the boolean method

Further Work

- Encode document importance in p(d|ca)
- Below the document level
- Lift the *position independence* assumption

Questions?

Krisztian Balog

kbalog@science.uva.nl
http://www.science.uva.nl/~kbalog