Associating People and Documents

Krisztian Balog and Maarten de Rijke

ISLA, University of Amsterdam
kbalog,mdr@science.uva.nl

Abstract. Since the introduction of the Enterprise Track at TREC in 2005, the task of finding experts has generated a lot of interest within the research community. Numerous models have been proposed that rank candidates by their level of expertise with respect to some topic. Common to all approaches is a component that estimates the strength of the association between a document and a person. Forming such associations, then, is a key ingredient in expertise search models. In this paper we introduce and compare a number of methods for building document-people associations. Moreover, we make underlying assumptions explicit, and examine two in detail: (i) independence of candidates, and (ii) frequency is an indication of strength. We show that our refined ways of estimating the strength of associations between people and documents leads to significant improvements over the state-of-the-art in the end-to-end expert finding task.

1 Introduction

Since the launch of the TREC Enterprise track [4, 10] there has been a lot of work on models, algorithms, and evaluation methodology for the expert finding task, i.e., the task of returning a list of people within some given organization that are ranked by their expertise on some given topic. A feature shared by many of the models proposed for ranking people with respect to their expertise on a given topic is their reliance on *associations* between people and documents. E.g., if someone is strongly associated with an important document on a given topic, this person is more likely to be an expert on the topic than someone who is not associated with any documents on the topic.

Despite the important role of associations between candidate experts (from now on: "candidates") and documents for today's expert finding models, such associations have received relatively little attention in the research community. Various methods have been used for estimating the strength of associations, and these approaches come in two kinds: (i) *set-based*, where the candidate is associated with a set of documents (all with equal weights), in which (s)he is mentioned, and (ii) *frequency-based*, where the strength of the association is proportional to the number of times the candidate is mentioned in the document.

While a number of techniques have already been used to estimate the strength of association between a person and a document, these have never been compared. This gives rise to the research questions that we seek to answer in this paper: 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 these associations? How sensitive are expert finding models to different document-candidate association methods?

To answer our research questions, we use two principal expert search strategies (so-called candidate and document models), that cover most existing approaches developed for expert finding. Our models are based on generative language modeling techniques, which is a specific choice, but the need for estimating the strength of the association between document-candidate pairs is not specific to our models. Other approaches also include this component, not necessarily in terms of probabilities, but as a score or weight. Given these models, we study, and systematically compare, various association methods. To this end we first discuss the *boolean* model, which is a simple yet effective way of forming associations, and serves as our baseline approach to person-document associations.

Then we lift an assumption that underlies this method—the *independence of candidates*—, and use term weighting schemes familiar from Information Retrieval. The strategy we follow is this: we treat candidates as terms and view the problem of estimating the strength of association with a document as an importance estimation problem: how important is a candidate for a given document. Specifically, we consider TF, IDF, TFIDF, and language models.

As a next step, we examine a second assumptions underlying (at least some) document-person association methods: that *frequency is an indication of strength*. First, we consider *lean* document representations that contain only candidates, while all other terms are filtered out. We find that it seriously impacts the performance of some expert-finding models (esp. candidate models) while it affects others to a far lesser degree (esp. document models).

Then, to grasp the effect of using the frequency of a candidate, we propose a new person-document association approach, where instead of the candidate's frequency, the *semantic relatedness* of the document and the person is used. This is achieved by comparing the language model of the document with the candidate's profile. We find that frequencies succeed very well at capturing the semantics of person-document associations.

The remainder of the paper is organized as follows. In Section 2 we discuss related work. We describe our expert search models in Section 3 and our experimental setup in Section 4. We compare multiple people-document association methods, and report our results in Section 5. We conclude in Section 6.

2 Related Work

In this section we focus on expertise retrieval approaches developed and published since the launch of the TREC Enterprise Track in 2005. For an overview of expertise finding systems in organizations we refer the reader to [12].

There are two principal approaches to expert finding, which have been first formalized and extensively compared in [1], and are referred to as *candidate* and *document* models. Most systems that took part in the 2005 and 2006 editions of the Expert Finding task at TREC implemented (variations on) one of them; see

[4, 10]. Candidate-based approaches (also referred to as profile-based methods in [5] or query-independent approaches in [9]) build a textual (usually term-based) representation (profile) of candidate experts, and rank them based on the relevance of a query, using traditional ad-hoc retrieval models [1, 5, 6, 7, 8, 9]. Document-based models (called query-dependent approaches in [9]) first locate documents on the topic and then find the associated experts [1, 2, 3, 5].

Common to all approaches is that documents and candidates need to be linked, whether these associations are made explicit or encoded in the models. Association methods come two kinds: (i) *set-based*, where the candidate is associated with a set of documents (all with equal weights), in which (s)he occurs [7, 8], and (ii) *frequency-based*, where the strength of the association is proportional to the number of times the candidate occurs in the document [1, 5, 6, 9].

In [7, 8] candidate profiles are constructed based on a set of documents in which the person's name or email address occurs. The candidate's identifier(s) (name and/or e-mail address) are used as a query, and relevant documents contribute to this set of profile documents. These approaches do not quantify the strength of the document-candidate associations, thus use them implicitly. In our setting this corresponds to the *boolean* model of associations (Section 5.1), i.e., a person is either associated with a document or not.

Document-based expert finding models often employ language models (LMs) [1, 2, 3, 5, 9] and the strength of the association between candidate ca and document d is expressed as a probability (either $p(d|ca)$ or $p(ca|d)$). In [1], these probabilities are calculated using association scores between document-candidate pairs. The scores are computed based on the recognition of the candidate's name and e-mail address in documents. In [5, 9], $p(d|ca)$ is rewritten in terms of $p(ca|d)$, using Bayes' rule, then the candidate's representations are treated as a query given the document model. This corresponds to our LM approach in Section 5.2. The two-stage LM [3, 2] includes a co-occurrence model, $p(ca|d, q)$, which is calculated based on the co-occurrence of the person with one or more query terms in the document or in the same window of text. When co-occurrence is calculated based on the full body of the document, the query is not taken into account and document-candidate associations are estimated using LMs, where documents contain only candidate identifiers. This corresponds to our lean documents approach using LMs in Section 5.3.

The candidate-generation model in [5] covers the two-stage LM approach of [3], but it is assumed that the query q and candidate ca are independent given the document d, i.e., $p(ca|d, q) \approx p(ca|d)$. The document model in [1] (Model 2 in Section 3) makes the same assumption. That implies that we build associations on the document level only, and leave an exploration of candidate-"text snippet" associations (co-occurrence on the sub-document level) for future work.

3 Modeling Expert Search

In this section we briefly describe two models for expert finding, taken from [1]. These two models cover both expert search strategies; moreover, they are principled, and nicely demonstrate how the document-association component fits into the picture. We should point out that these models consider document-candidate associations on the document level only.

3.1 Model 1: Candidate Model

Model 1 builds a textual representation of a candidate ca using a multinomial unigram language model θ_{ca}. This model is used to predict how likely a candidate would produce a query q:

$$p(q|\theta_{ca}) = \prod_{t \in q} p(t|\theta_{ca})^{n(t,q)} \tag{1}$$

where each term t in the query q is sampled identically and independently, and $n(t,q)$ is the number of times t occurs in q.

The candidate model is constructed as a linear interpolation of an empirical model $p(t|ca)$, and the background model $p(t)$ to ensure there are no zero probabilities:

$$p(t|\theta_{ca}) = (1-\lambda) \cdot p(t|ca) + \lambda \cdot p(t), \tag{2}$$

where parameter λ controls the amount of smoothing applied.

Using the associations between a candidate and a document, the probability $p(t|ca)$ can be approximated by $p(t|ca) = \sum_d p(t|d) \cdot p(d|ca)$, where $p(d|ca)$ is the probability that candidate ca generates supporting document d, and $p(t|d)$ is the probability of term t occurring in document d (calculated using the maximum-likelihood estimate of a term). The final estimation of Model 1 is:

$$p(q|\theta_{ca}) = \prod_{t \in q} \left\{ (1-\lambda) \cdot \left(\sum_d p(t|d) \cdot p(d|ca) \right) + \lambda \cdot p(t) \right\}^{n(t,q)} \tag{3}$$

3.2 Model 2: Document Model

Under this model, we can think of the process of finding an expert as follows. Given a collection of documents ranked according to the query, we examine each document and if relevant to our problem, we then see who is associated with that document. Conceptually, Model 2 differs from Model 1 because the candidate is not directly modeled. Instead, it assumes that q and ca are independent given d, the document acts like a "hidden" variable in the process which separates the query from the candidate. Formally, this can be expressed as

$$p(q|ca) = \sum_d p(q|d) \cdot p(d|ca) \tag{4}$$

The probability of a query given a document $p(q|d)$ is estimated by inferring a document language model θ_d for each document d:

$$p(t|\theta_d) = (1-\lambda) \cdot p(t|d) + \lambda \cdot p(t) \tag{5}$$

where $p(t|d)$ is the probability of the term in the document. The probability of a query given the document model is:

$$p(q|\theta_d) = \prod_{t \in q} p(t|\theta_d)^{n(t,q)} \tag{6}$$

The final estimate of Model 2, then, is:

$$p(q|ca) = \sum_d \left\{ \prod_{t \in q} \left((1-\lambda) \cdot p(t|d) + \lambda \cdot p(t) \right)^{n(t,q)} \right\} \cdot p(d|ca) \tag{7}$$

3.3 Document-Candidate Associations

In Model 1 and 2 the association between candidate ca and document d is expressed as $p(d|ca)$, the probability of the document given the candidate. We apply Bayes' rule to rewrite it:

$$p(d|ca) = \frac{p(ca|d) \cdot p(d)}{p(ca)} \tag{8}$$

This allows us to incorporate document and candidate priors into the association component. We leave the estimation of document and candidate priors to future work and assume that $p(d)$ and $p(ca)$ are uniformly distributed. Hence, our task boils down to estimating of $p(ca|d)$. The reading of $p(ca|d)$ is different for the two models. For Model 1, it reflects the degree to which the candidate's expertise is described using this document. For Model 2, it provides a ranking of candidates associated with a given document d, based on their contribution made to d.

4 Experimental Setup

4.1 Test Collection

We use the test sets of the 2005 and 2006 editions of the TREC Enterprise track [4, 10]. The document collection used is the W3C corpus [11], a heterogenous document repository containing a mixture of document types crawled from the W3C site. We used the entire corpus, and handled all documents in the same way, as HTML documents. We did not resort to any special treatment of document types, nor did we exploit the internal document structure that may be present; instead, we represented all documents as plain text. We removed a standard list of stopwords, but did not apply stemming.

The TREC Enterprise 2005 topics (50) are names of working groups within the W3C. Members of a working group were regarded as experts on the corresponding topic. The 2006 topics (55) were contributed by TREC participants and assessed manually. We used only the titles of the topic descriptions.

We evaluate the methods with mean average precision (MAP), the official measure of the expert finding task at TREC.

4.2 Person Name Identification

In order to form document-candidate associations, we need to be able to recognize candidates' occurrences within documents. In the TREC setting, a list of possible candidates is given, where each person is described with a unique *person_id*, one or more *names*, and one or more *e-mail* addresses. While this is a specific way of identifying a person, and different choices are also possible (e.g., involving social security number instead of, or in addition to, the representations just listed), nothing in our modeling depends on *this* particular choice.

The recognition of candidate occurrences in documents (through one of these representations) is a restricted information extraction task. In [2], six match types (MT) of person occurrences are identified:

MT1 Full name (e.g., Ritu Raj Tiwari and Tiwari, Ritu Raj);

MT2 Email name (e.g., rtiwari@nuance.com);
MT3 Combined name (e.g., Tiwari, Ritu R and R R Tiwari);
MT4 Abbreviated name (e.g., Ritu Raj and Ritu);
MT5 Short name (e.g., RRT);
MT6 Alias, New Mail (e.g., Ritiwari and rtiwari@hotmail.com).

In [1], a similar approach is taken, and four types of matching are introduced; three attempt to identify candidates by name, and one uses email addresses. To facilitate comparison, we used the resources contributed by Bao et al. [2].[1]

Some of these matching methods create ambiguity, that is, a name may be shared by more than one person. To allow us to measure, how this noise introduced affects overall performance, we identify a group of matching methods, including `MT1`, `MT2`, and `MT6`, where ambiguity is insignificant, and refer to this set as `STRICT` matching methods. Using all matching methods is referred as `ALL`.

We replaced all candidate occurrences (name and email address) with a unique candidate identifier, which was then treated as a term in the document.

5 Establishing Document-Candidate Associations

In this section we address the problem of estimating $p(ca|d)$, the strength of the association between a document and a candidate.

5.1 The Boolean Model of Associations

Under the boolean model, associations are binary decisions; they exist if the candidate occurs in the document, irrespective of the number of times the person or other candidates are mentioned in that document. We simply set

$$p(ca|d) = \begin{cases} 1, n(ca,d) > 0 \\ 0, \text{otherwise}, \end{cases} \tag{9}$$

where $n(ca, d)$ denotes the number of times the candidate's identifier appears in the document. It can be viewed as a set-based approach, analogously to [7], where a candidate is associated with a set of documents: $D_{ca} = \{d : n(ca, d) > 0\}$.

The boolean model is the simplest way of forming document-candidate associations. Simplicity comes at the price of two potentially unrealistic assumptions:

- **Candidate independence** Candidates occurring in the document are independent of each other, and are all equally important given the document. The model does not differentiate between people that occur in its text.
- **Position independence** The strength of the association between a candidate and a document is independent of the candidate's position within the document. Positional independence is equivalent to adopting the bag of words representation: the exact ordering of candidates within a document is ignored, only the number of occurrences is stored.

[1] URL: http://ir.nist.gov/w3c/contrib/

Common sense tells us that not all candidates mentioned in the document are equally important. Similarly, not all documents, in which a candidate occurs, describe the person's expertise equally well. For example, a person who is listed as an author of the document should be more strongly associated with the document, than someone who is only referred to in the body of the document. This goes against the candidate independence assumption. If we take into account that authors are also listed at the top or bottom of documents, the previous example also provides evidence against the position independence assumption.

In this paper, we stick with the position independence assumption, and leave the examination of that to further work. However, intuitively, candidate independence may be too strong an assumption. Therefore, we drop it as our next step, and discuss ways of estimating a candidate's importance given a document. In other words, our aim is a non-binary estimation of $p(ca|d)$.

5.2 Modeling Candidate Frequencies

Our goal is to formulate $p(ca|d)$ in such a way that it indicates the strength of the association between candidate ca and document d. The number of times a person occurs in a document seems to be the most natural evidence supporting the candidate being strongly associated with that document. This leads us to a new assumption: the strength of the association is proportional to the number of times the candidate is mentioned in the document.

A commonly employed technique for building document-candidate associations is to use the candidate's identifiers as a query to retrieve documents. The strength of the association is then estimated using the documents' relevance scores [5, 9]. This way, both the recognition of candidates' occurrences and the association's strength estimation is performed in one step. Our approach is similar, but limited to the estimation aspect, and assumes that the matching of candidate occurrences is taken care of by a separate extraction component.

We treat candidate identifiers as terms in a document, and view the problem of estimating the strength of association with a document as an importance estimation problem: how important is a candidate for a given document? We approach it by using term weighting schemes familiar from IR. Specifically, we consider TF, IDF, and TFIDF weighting schemes from the vector space model, and also language models. In the following, we briefly discuss these methods and the rationale behind them.

TF The importance of the candidate within the particular document is proportional to the candidate's frequency (against all terms in the document): $p(ca|d) \propto TF(ca, d)$

IDF It models the general importance of a candidate:

$$p(ca|d) \propto \begin{cases} IDF(ca), & n(ca, d) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Candidates that are mentioned in many documents, will receive lower values, while those who occur only in a handful of documents will be compensated with higher values. This, however is independent of the document itself.

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

	ALL MatchTypes				STRICT MatchTypes			
Method	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	$.0642^{(3)}$	$.2031^{(2)}$	$.1586^{(3)}$	$.4470^{(1)}$

TFIDF A combination of the candidate's importance within the particular document, and in general is expected to give the best results.

Language Modeling We employ a standard LM setting to document retrieval, using Equation 5. We set $p(ca|d) = p(t = ca|\theta_d)$, which is identical to the approach in [5, 9]. Our motivation for using language models is twofold: (i) expert finding models are also using LMs (pragmatic reason), and more importantly, and (ii) smoothing in language modeling has an IDF effect [13]. Tuning the value of λ allows us to control the background effect (general importance of the candidate), which is not possible using TFIDF. Here, we follow standard settings and use $\lambda = 0.1$ [13].

Table 1 presents the MAP scores for Model 1 and 2, using the TREC 2005 and 2006 topics. We report on two sets of experiments, using all (columns 2–5) and only the unambiguous (columns 6–9) matching methods. The first row corresponds to the boolean model of associations (Eq. 10), while additional rows correspond to frequency-based methods.

For significance testing we use a two-tailed, matched pairs Student's t-test, and look for improvements at significance levels $^{(1)}$ 0.95, $^{(2)}$ 0.99, and $^{(3)}$ 0.999. The boolean method is considered as the baseline, against which frequency-based methods are compared.

Our findings are as follows. First, there is a substantial difference between the performance on the TREC 2005 and 2006 topic sets. As pointed out in [5], this is due to the fact that judgments were made differently in these two years. In 2005, judgments are independent of the document collection, and were obtained artificially, while topics in 2006 were developed and assessed manually. Second, it is more beneficial to use rigid patterns for person name matching; the noise introduced by name ambiguity hurts performance. Hence, from now on we use the STRICT matching methods. Third, Model 2 performs considerably better than Model 1. This confirms the findings reported in [1].

As to the association methods, we find that the simple boolean model delivers excellent performance. The best results (using Model 2 and STRICT matching) are 0.2196 and 0.4704 for TREC 2005 and 2006, respectively; this beats the corresponding scores of 0.204 and 0.465 scores of Fang and Zhai [5]. However, in

[5] candidate priors are used, and parameters of the models are tuned, while we use baseline settings. Compared with the official results of the TREC Enterprise track [4, 10], results produced by our boolean model would be in the top 3 for 2005 and top 10 for 2006. Top performing systems tend to use various kinds of heuristics, manual topic expansion, and sub-document models.

Surprisingly, in most cases the boolean model performed better than the frequency-based weighting schemes. The only noticeable difference is for Model 2 using the 2005 topics, where the IDF weighting achieves up to 0.25 MAP. The explanation of this behavior, again, lies in the nature of the 2005 topic set. Relevant experts in TREC 2006 are more popular in the collection compared to those identified in TREC 2005 [5], which means that penalizing popular candidates, which is indeed what IDF does, is beneficial for TREC 2005. Importantly, Model 1 shows much more variance in accuracy than Model 2. In case of the more realistic 2006 topic set, the use of various methods for Model 2 indicate hardly any difference. To explain this effect, we need to consider the inner workings of these two strategies. In case of the candidate model (Model 1), document-candidate associations determine the degree to which a document contributes to the person's profile. If the candidate is a "regular term" in the document, shorter documents contribute more to the profile than longer ones. E.g., if the person is an author of a document and appears only at the top of the page, a shorter document influences her profile more than a longer one. Intuitively, a length normalization effect would be desired to account for this. The boolean approach adds all documents with the same weight to the profile, and as such, does not suffer from this effect. On the other hand, this simplification may be inaccurate, since all documents are handled as if authored by the candidate.

For the document model (Model 2), we can observe the same length normalization effect. E.g., if two documents d_1, d_2 contain the same candidates, but have $|d_1| = 1000$ and $|d_2| = 250$, while the relevance scores of these documents are 1 and 0.5, respectively, then d_2 will add twice as much as d_1 to the final expertise score, even though its relevance is lower.

5.3 Using Lean Documents

To overcome the length normalization problem, we propose a *lean document representation*, where documents contain only candidate identifiers, and all other terms are filtered out. This can be viewed as "extreme stopwording," where all terms except candidate identifiers are stopwords. Given this representation, the same weighting schemes are used as before. Calculating TF on lean documents is identical to the candidate-centric way of forming associations proposed in [1]. IDF values remain the same, as they rely only on the number of documents in which the candidate occurs, which is unchanged.

For language models, the association's strength is calculated using

$$p(ca|d) = (1 - \lambda) \cdot \frac{n(ca, d)}{|d|} + \lambda \cdot \frac{n(ca)}{\sum_{d'} |d'|}, \tag{11}$$

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%)

where $|d|$ denotes the length of d (total number of candidate occurrences in d), and $n(ca) = \sum_{d'} n(ca, d')$. Essentially, this is the same as the so-called document-based co-occurrence model of Cao et al. [3].

Table 2 presents the results. Significance is tested against the normal document representation (corresponding rows of Table 1, STRICT MatchTypes). The numbers in brackets denote the relative changes in performance.

For Model 1, using the lean document representation shows improvements of up to 227% compared to the standard document representation, and up to 24% compared to the boolean approach (differences are statistically significant). This shows the need of the length normalization effect for candidate-based approaches, such as Model 1, and makes frequency-based weighting schemes using lean documents a preferred alternative over the boolean method.

As to Model 2, the results are mixed. Using the lean document representation instead of the standard one hurts for the TREC 2005 topics, and shows moderate improvement (up to 4.7%) on the 2006 topics. For the document-based expert retrieval strategy the relative ranking of candidates for a fixed document is unchanged, and the length normalization effect is apparently of less importance than for the candidate-based model. Compared to the boolean association method, there is no significant improvement in performance (except the IDF weighting for 2005, which we have discussed earlier).

5.4 Semantic Relatedness

So far, we have used the number of times a candidate occurs in a document as an indication of its importance for the document. We will now re-visit this assumption. We propose an alternative way of measuring the candidate's weight in the document—semantic relatedness. We use the lean document representation, but a candidate is represented by its semantic relatedness to the given document, instead of its actual frequency. We use $n'(ca, d)$ instead of $n(ca, d)$, where

$$n'(ca, d) = \begin{cases} \mathrm{KLDIV}(\theta_{ca}||\theta_d), & n(ca, d) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

That is, if the candidate is mentioned in the document, his weight will be the distance between the candidate's and the document's language models, where the document's language model is calculated using Eq. 5 and the candidate's language model is calculated using Model 1, Eq. 3.

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 absolute performance of the association method based on semantic relatedness is in the same general range as the frequency-based association method listed alongside it. Columns 4, 7, 10, 13 provide the Kendall tau rank correlation scores for the two columns that precede them—which are very high indeed. These correlation scores suggest that frequency-based associations based on lean documents are capable of capturing the semantics of the associations.

6 Discussion and Conclusions

As a retrieval task, expert finding has attracted much attention since the launch of the Enterprise Track at TREC in 2005. Two clusters of methods emerged, so-called candidate and document models. Common to these approaches is a component that estimates the strength of the association between a document and a person. Forming such associations is a key ingredient, yet this aspect has not been addressed as a research topic. In this paper we introduced and systematically compared a number of methods for building document-people associations. We made explicit a number of assumptions underlying various association methods and analyzed two of them in detail: (i) independency of candidates, and (ii) frequency is an indication of strength.

We gained insights in the inner workings of the two main expert search strategies, and found that these behave quite differently with respect to document-people associations. Candidate-based models are sensitive to associations. Lifting the candidate independence assumption and moving from boolean to frequency based methods can improve performance by up to 24%. However, the standard document representation (where candidate occurrences are treated as regular terms) suffers from length normalization problems, therefore, a lean document representation (that contains only candidates, while all other terms are filtered out) should be used.

On the other hand document-based models are less dependent on associations, and the boolean model turned out to be a very strong baseline. Only a moderate (up to 4.7%) improvement can be gained by moving to frequency-based associations. Absolute scores of the document-based model are substantially higher than of the candidate-based one, which makes it the preferred strategy.

To assess the *frequency is an indication of strength* assumption we proposed a new people-document association approach, based on the semantic relatedness of the document and the person. We find that frequencies succeed very well at capturing the semantics of person-document associations.

This study suggest that this is how far we can get by capturing expertise at the document level. For further improvements we seem to need sub-document models as well corpus-specific methods but in a non-heuristic way.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (NWO) by grants under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.066.302, 612.069.006, 640.001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

Bibliography

[1] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06*, pages 43–50, New York, NY, USA, 2006.

[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 *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, 2007.

[3] Y. Cao, J. Liu, S. Bao, and H. Li. Research on Expert Search at Enterprise Track of TREC 2005. In *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, 2006.

[4] N. Craswell, A. de Vries, and I. Soboroff. Overview of the TREC-2005 Enterprise Track. In *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, 2006.

[5] H. Fang and C. Zhai. Probabilistic models for expert finding. In *Proceedings ECIR 2007*, pages 418–430, 2007.

[6] Y. Fu, W. Yu, Y. Li, Y. Liu, and M. Zhang. THUIR at TREC 2005: Enterprise Track. In *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, 2006.

[7] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM '06*, pages 387–396, 2006.

[8] C. Macdonald, V. Plachouras, B. He, and I. Ounis. University of Glasgow at TREC2005: Experiments in Terabyte and Enterprise tracks with Terrier. In *Proceedings of the 14th Text REtrieval Conference (TREC 2005)*, 2005.

[9] D. Petkova and W. B. Croft. Hierarchical language models for expert finding in enterprise corpora. In *ICTAI'06*, pages 599–608, 2006.

[10] I. Soboroff, A. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise Track. In *TREC 2006 Working Notes*, 2006.

[11] W3C. The W3C test collection, 2005. URL: http://research.microsoft.com/users/nickcr/w3c-summary.html.

[12] D. Yimam-Seid and A. Kobsa. Expert finding systems for organizations: Problem and domain analysis and the demoir approach. *Journal of Organizational Computing and Electronic Commerce*, 13(1):1–24, 2003.

[13] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, pages 334–342, 2001.