Non-Local Evidence for Expert Finding

Krisztian Balog
kbalog@science.uva.nl

Maarten de Rijke
mdr@science.uva.nl

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam

ABSTRACT

The task addressed in this paper, finding experts in an enterprise setting, has gained in importance and interest over the past few years. Commonly, this task is approached as an association finding exercise between people and topics. Existing techniques use either documents (as a whole) or proximity-based techniques to represent candidate experts. Proximity-based techniques have shown clear precision-enhancing benefits. We complement both document and proximity-based approaches to expert finding by importing global evidence of expertise, i.e., evidence obtained using information that is not available in the immediate proximity of a candidate expert's name occurrence or even on the same page on which the name occurs. Examples include candidate priors, query models, as well as other documents a candidate expert is associated with.

Using the CERC data set created for the TREC 2007 Enterprise track we identify examples of non-local evidence of expertise. We then propose modified expert retrieval models that are capable of incorporating both local (either document or snippet-based) evidence and non-local evidence of expertise. Results show that our refined models significantly outperform existing state-of-the-art approaches.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3 .3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [**Information Systems Applications**]: H.4.2 Types of Systems; H.4.m Miscellaneous

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Expert finding, enterprise search, query models, language models, priors

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.

1. INTRODUCTION

Expert finding addresses the task of finding the right person with the appropriate skills and knowledge [5]. Expert finding systems rank candidate experts with respect to a given topic. A key ingredient of such systems is to compute associations between candidates and topics that capture how strong the two are related. Usually, such associations are determined by considering the documents in which candidates and topics co-occur and more recently such associations have been computed not at the document-level but more locally, using text windows or snippets around occurrences of names of candidate experts.

On the whole, it has been found that the use of local, proximity-based evidence for computing associations between candidate experts and topics improves precision on the overall expert finding task. This is not a surprise. However, there are additional types of evidence of a candidate's expertise in a given topical area that are distinctively non-local in character. By *non-local* evidence we mean evidence of expertise that is not available from an individual text snippet or even from an individual page. To make matters concrete we provide a number of examples. We take these examples from the experimental setting provided by the TREC 2007 Enterprise track [2] and its scenario of science communicators in a knowledge intensive organization (CSIRO, [1]) that have to recommend experts in response to outside requests for experts; despite this specific choice, we believe that the phenomenon of non-local indicators of expertise is completely general and generic.

One type of non-local evidence relates to clickstream data; if we have seen the topic for which expertise is being sought before, say in a document retrieval setting, and we have examples of key documents that are often clicked on, how can we use this information about the topic to improve the discovery of associations between candidate experts and topics?

Another type of non-local evidence concerns the (relative) importance of a candidate for a given document ($(p(ca|d))$). A candidate expert that is related to many documents may not have been particularly important for the creation of a given document d; thus, in turn, d probably should not contribute a lot as evidence in support of associations between the candidate ca and topics discussed in d. And similarly, if the documents associated with a candidate are not semantically related to a given document d, then, again, this particular document probably should not count heavily as evidence in support of associations between the candidate ca and topics discussed in d.

A final example suggests that we should consider global

properties of candidate experts when computing expert-topic associations: the mere co-occurrence of a person with a topic need not be an indication of expertise of that person on the topic. A case in point is provided by the science communicators in the CSIRO enterprise search test set: they are mentioned as a contact person on many pages and, hence, frequently co-occur with many topics.

Our aim in this paper is to identify, model, and estimate non-local sources of evidence for expert finding and to integrate such evidence into existing language modeling-based approaches to expert finding. Concretely, we aim to find out to which extent rich query modeling with non-local evidence improves the effectiveness of expert finding. Second, we seek to determine how different ways of computing document-expert associations (with different types of global statistics) impact expert finding. And, we explore to which extent priors on candidate experts (based on their global co-occurrence behavior) impact expert finding effectiveness.

Our main contributions are a comparison of existing expert search approaches on the TREC 2007 enterprise platform (CERC collection), the identification of a number of non-local sources of expert finding as well as ways of estimating and modeling these in an effective way, based on existing document and proximity-based approaches to expert finding.

The remainder of the paper is organized as follows. We discuss related work in Section 2. We detail our models and ways of estimating both local and non-local evidence for expertise in Section 3. Our experimental setup is detailed in Section 4 and we report on our experiments in Section 5. Section 6 contains an analysis of our experimental results and we conclude in Section 7.

2. RELATED WORK

To reflect the growing interest in entity ranking in general and expert finding in particular, TREC introduced an expert finding task at its Enterprise track in 2005 [11]. At this track it emerged that there are two principal approaches to expert finding—or rather, to capturing the association between a candidate expert and an area of expertise [11, 26, 2]. The two models have been first formalized and extensively compared by Balog et al. [5], and are called *candidate* and *document* models, or *Model 1* and *Model 2*, respectively. Model 1's candidate-based approach is also referred to as profile-based method in [12] or query-independent approach in [19]. These approaches build a textual (usually term-based) representation of candidate experts, and rank them based on query/topic, using traditional ad-hoc retrieval models. Conceptually, these approaches are similar to the P@noptic system [10]. The other type of approach, document models, are also referred to as query-dependent approaches in [19]. Here, the idea is to first find documents which are relevant to the topic, and then locate the associated experts. Thus, Model 2 attempts to mimic the process one might undertake to find experts using a document retrieval system. Nearly all systems that took part in the 2005–2007 editions of the Expert Finding task at TREC implemented (variations on) one of these two approaches. In this paper we focus on (variations on) Model 1.

Building on either candidate or document models, further refinements to estimating the association of a candidate with the topic of expertise have been explored. For example, instead of capturing the associations at the document level, they may be estimated at the paragraph or snippet level [3]. The generative probabilistic framework naturally lends itself to such extensions, and to the inclusion of other forms of evidence, such as document and candidate evidence through the use of priors [12], document structure [28], and of hierarchical, organizational and topical context and structure [19, 6]. For example, Petkova and Croft [19] propose another extension to the framework, where they explicitly model the topic, in a manner similar to relevance models for document retrieval [13]. The topic model is created using pseudo-relevance feedback, and is matched against document and candidate models. Serdyukov and Hiemstra [23] propose a person-centric method that combines the features of both document- and profile-centric expert finding approaches.

Fang and Zhai [12] demonstrate how query/topic expansion techniques can be used within the framework; the authors also show how the two families of models (i.e., Model 1 and 2) can be derived from a more general probabilistic framework. Petkova and Croft [20] introduce effective formal methods for explicitly modeling the dependency between the named entities and terms which appear in the document. They propose candidate-centered document representations using positional information, and estimate $p(t|ca, d)$ using proximity kernels. Their approach is similar to the window-based models that we use below, in particular, their step function kernel corresponds to our estimate of $p(t|ca, d)$ in Eq. 10 below. Balog and de Rijke [4] introduce and compare a number of methods for building document-candidate associations. Empirically, the results produced by such models have been shown to deliver state of the art performance (see [5, 19, 20, 12, 6]).

Finally, we highlight two alternative approaches that do not fall into the categories above (i.e., candidate or document models). Macdonald and Ounis [15] propose ranking experts with respect to a topic based on data fusion techniques, without using collection-specific heuristics; they find that applying field-based weighting models improves the ranking of candidates. Macdonald et al. [17] enhance their voting approach by considering proximity, moreover, experiment with integrating additional evidence by identifying home pages of candidate experts and clustering relevant documents. The authors report experimental results on the TREC 2007 platform (CERC) in [14, 17]. Rode et al. [22] represent documents, candidates, and associations between them as an entity containment graph, and propose relevance propagation models on this graph for ranking experts.

Independent of the basic model adopted, various teams have worked on improved query modeling in the setting of expert finding and, more generally, enterprise search. E.g., Macdonald and Ounis [16] studied better query modeling with query expansion for expert finding and Balog et al. [7] explored query expansion in the setting of enterprise search using so-called example documents (sample key pages are provided with the topic description; see the description of "feedback runs" below). In some of the manual runs produced at TREC 2007 improved query modeling was obtained by manually tuning queries derived from the narrative field of the topic statements [29].

In Table 1 we list the highest scoring results achieved using the TREC 2007 test set (CERC, [1]) that we have been able to find in the literature. We distinguish between three types of runs: *automatic*, *feedback*, and *manual*. Manual runs involve humans in the loop at any stage, for example

Method/model	MAP	P5	P10	MRR
TREC 2007 best	.4632	.2280		
TREC 2007 best feedback	.3660	.2040		
TREC 2007 best manual	.4787	.2720		
Voting model [14]	.3406		.1224	
Voting model [17]	.3519			.4730
Voting model+proximity [17]	.4319			.5742
Relevance prop. [24]	.4528			.5840
Model 1 [3]	.3801	.2000	.1340	.5571
Model 2 [3]	.4142	.2400	.1620	.5671
Model 1B [3]	.4633	.2600	.1620	.6236
Model 2B [3]	.4323	.2560	.1600	.5790

Table 1: Numbers reported so far in the literature on the TREC 2007 Enterprise platform.

composing queries from the topics, manual term expansion, relevance feedback, or manual combination of results. Feedback runs can be thought of as simulating one type of click-based system. They involve the use of the title and page fields of the topics in the TREC 2007 topic set; the page field contains examples of key reference URLs (on average 4 per topic)—these simulate the situation where we have seen the query before (in a document retrieval setting) and a few URLs were often clicked: the URLs in the pages field.

The first group of results in Table 1 are the highest scoring runs at TREC 2007 [2]. The second group is produced using Macdonald and Ounis's fusion techniques. The third group represents the best scores obtained using the graph-based approach of [24]. The fourth and fifth group represent the original candidate and document models and their window-based refinements, respectively.

It was found, both at TREC 2007 and afterwards, that performance depends on two critical factors: the ability to accurately recognize name occurrences in document[1] and the choice of parameters: wherever possible, we use the best or optimal parameter settings as reported in the literature.

3. MODELING

Within an organization, there may be many possible candidates who could be experts on a given topic. For a given query, the problem is to identify which of these candidates are likely to be an expert. Following [5] we can state this problem as follows:

> what is the probability of a candidate ca being an expert given the query topic q?

That is, we wish to determine $p(ca|q)$, and rank candidates ca according to this probability. The candidates with the highest probability given the query are deemed to be the most likely experts for that topic. The challenge, of course, is how to accurately estimate this probability. Instead of calculating this probability directly we apply Bayes' rule and rewrite it to

$$p(ca|q) = \frac{p(q|ca) \cdot p(ca)}{p(q)}, \tag{1}$$

where $p(ca)$ is the probability of a candidate and $p(q)$ is the probability of a query. Since $p(q)$ is a constant (for a given query), it can be ignored for the purpose of ranking. Thus, the probability of a candidate ca being an expert given the query q is proportional to the probability of a query given the candidate $p(q|ca)$, weighted by the *a priori* belief that candidate ca is an expert ($p(ca)$):

$$p(ca|q) \propto p(q|ca) \cdot p(ca). \tag{2}$$

In most existing work [5, 9] $p(ca)$ is assumed to be uniform. However, as was shown in [12], a reasonable prior can improve retrieval accuracy. In this paper we will use candidate priors to distinguish between science communicators and proper experts; the estimation of this prior is detailed in Section 3.4.

According to Model 1 of Balog et al. [5], the candidate is represented by a multinomial probability distribution over the vocabulary of terms. Therefore, a candidate model θ_{ca} is inferred for each candidate ca, such that the probability of a term given the candidate model is $p(t|\theta_{ca})$. The model is then used to predict how likely a candidate would produce a query q.

Assuming that each query term is sampled identically and independently, the query likelihood is obtained by taking the product across all the terms in the query, such that:

$$p(q|\theta_{ca}) = \prod_{t \in q} p(t|\theta_{ca})^{n(t,q)}, \tag{3}$$

where $n(t, q)$ denotes the number of times term t is present in query q.

Instead of calculating this probability directly, we move to the log domain to prevent numerical underflows, as proposed in [3]. We rewrite Eq. 3 as follows:

$$\log p(q|\theta_{ca}) = \sum_{t \in q} p(t|\theta_q) \cdot \log p(t|\theta_{ca}). \tag{4}$$

In this alternative formulation we also replaced $n(t, q)$ with $p(t|\theta_q)$, which can be interpreted as the weight of term t in query q. We will refer to θ_q as the *query model*. Note that maximizing the query-likelihood in Eq. 4 provides the same ranking as minimizing the KL-divergence between the query and candidate models (that is, ranking by $-\mathrm{KL}(\theta_q||\theta_{ca})$) (as is pointed out in [7]).

Next, we discuss the estimation of the three components of our modeling: (i) the candidate model ($p(t|\theta_{ca})$) in Section 3.1, (ii) the query model ($p(t|\theta_q)$) in Section 3.3, and (iii) candidate priors ($p(ca)$) in Section 3.4. Along the way, in Section 3.2, we discuss a key ingredient of our candidate models, viz. document-candidate associations ($p(ca|d)$).

3.1 Candidate Model

To obtain an estimate of $p(t|\theta_{ca})$, we must ensure that there are no zero probabilities due to data sparsity. In document language modeling, it is standard to employ smoothing:

$$p(t|\theta_{ca}) = (1 - \lambda_{ca}) \cdot p(t|ca) + \lambda_{ca} \cdot p(t), \tag{5}$$

where $p(t|ca)$ is the probability of a term given a candidate, and $p(t)$ is the probability of a term in the document repository.

To approximate $p(t|ca)$, we use the documents as a bridge to connect the term t and candidate ca in the following way:

$$p(t|ca) = \sum_{d} p(t, ca|d). \tag{6}$$

[1]To facilitate comparison we release a list of 3,490 names along with the documents associated with them at `http://es.csiro.au/cerc/data/balog`; the list is available for registered licensees of the CERC collection.

That is, the probability of selecting a term given a candidate is based on the strength of the co-occurrence between a term and a candidate in a particular document ($p(t, ca|d)$). Below, we first discuss two ways of building candidate models: based on documents associated with them and based on terms in proximity to candidate name mentions.

3.1.1 Document-based Model: Model 1

Our first approach to estimating candidate models assumes that the document and the candidate are conditionally independent. That is

$$p(t, ca|d) = p(t|d) \cdot p(ca|d), \tag{7}$$

where $p(t|d)$ is the probability of the term t in document d. We approximate it with the standard maximum-likelihood estimate of the term, i.e., the relative frequency of the term in the document [5].

3.1.2 Proximity-based Model: Model 1B

Model 1 assumes conditional independence between the document and the candidate. However, this assumption is quite strong as it suggests that all the evidence within the document is descriptive of the candidate's expertise. This may be the case if the candidate is the author of the document, but here we consider an alternative. We can factor the conditional probability $p(t, ca|d)$ as follows:

$$p(t, ca|d) = p(t|d, ca) \cdot p(ca|d). \tag{8}$$

That is, we base $p(t, ca|d)$ on the strength of the co-occurrence between a term and a candidate in a particular document; both the document and the candidate determine the probability of the term.

One natural way in which to estimate the probability of co-occurrence between a term and a candidate is by considering the proximity of the term given the candidate in the document, the idea being that the closer a candidate is to a term the more likely that term is associated with their expertise.

Here, we assume that the candidate's name, email, etc. have been replaced within the document representation with a unique candidate identifier, which can be treated much like a term, referred to as ca. The terms surrounding either side of ca form the context of the candidate's expertise and can be defined by a window of size w within the document. For any particular distance (window size) w between a term t and candidate ca, we can define the probability of a term given the candidate, distance and document:

$$p(t|ca, w, d) = \frac{n(t, ca, w, d)}{\sum_{t'} n(t', ca, w, d)}, \tag{9}$$

where $n(t, ca, w, d)$ is the number of times the term t co-occurs with ca at a distance of at most w in document d. Now, the probability of a term given the candidate and document is estimated by taking the sum over all possible window sizes W:

$$p(t|d, ca) = \sum_{w \in W} p(t|ca, w, d) \cdot p(w), \tag{10}$$

where $p(w)$ is the prior probability that defines the strength of association between the term and the candidate at distance w, such that $\sum_{w \in W} p(w) = 1$. Estimating Model 1B this way essentially corresponds to the step function kernel in [20].

When we put together our choices so far, the formula we use for ranking candidates is the one shown in Eq. 11.

$$p(ca|q) \propto p(ca) \cdot \sum_{t \in q} p(t|\theta_q) \cdot \log \Big\{ (1 - \lambda) \sum_{d} p(t, ca|d) + \lambda p(t) \Big\}. \tag{11}$$

So far we have discussed the estimation of $p(t, ca|d)$ and $p(t)$. Next, we discuss three additional components of the model: (i) document-candidate associations ($p(d|ca)$) in Section 3.2, (ii) the query model ($p(t|\theta_q)$) in Section 3.3, and finally, (iii) candidate priors ($p(ca)$) in Section 3.4.

3.2 Document-Candidate Associations

A feature common to both models introduced above, and shared by many of the models mentioned in Section 2, 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 or only with marginally relevant documents. In our framework this component is referred to as *document-candidate associations*, and the likelihood of candidate ca being associated with document d is expressed as a probability ($p(ca|d)$) in Eq. 7 for Model 1 and in Eq. 8 for Model 1B.

The probability $p(ca|d)$ can be estimated at the level of the document d itself, or at the sub-document level, where associations link people to specific text segments. To remain focused, we build associations on the document level only in this section: to date, many open issues remain even at the document level. We leave a systematic exploration of candidate-"text snippet" associations for later research.

We assume that the recognition of candidate occurrences is taken care of by an external extraction component. We briefly discuss this process in technical terms in Section 4. The output of this extraction procedure is a preprocessed document format where candidate occurrences are treated as terms. The number of times the candidate ca is recognized in the document d is denoted by $n(ca, d)$.

We take a baseline approach to computing $p(ca|d)$ together the two best performing approaches as suggested by Balog and de Rijke [4]. The simplest possible way of setting $p(ca|d)$ is referred as the *boolean model* (BL). Under this 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. Formally, it is expressed as:

$$p(ca|d) = \begin{cases} 1, & n(ca, d) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

For a better estimate, a lean document representation is used which consists of only candidate mentions. First the candidate's (local) frequency in the document (TF) and (global) frequency (IDF) is combined (and referred as TFIDF) (note that it is computed only for candidates that occur in the document ($n(ca, d) > 0$):

$$p(ca|d) \propto \frac{n(ca, d)}{\sum_{ca'} n(ca', d)} \cdot \log \frac{|D|}{|\{d' : n(ca, d') > 0\}|} \tag{13}$$

Note that this is a clear example of the use of non-local information (as we need global statistics to determine IDF).

Finally, we use an alternative way of measuring a candidate's importance given a document. A candidate is represented by its semantic relatedness to the given document, instead of its actual frequency. This method will be referred to as SEM. We use $n'(ca, d)$ instead of $n(ca, d)$ in Eq. 13, where

$$n'(ca, d) = \begin{cases} \mathrm{KL}(\theta_{ca}||\theta_d), & n(ca, d) > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

Again, we need global statistics to compute this way of determine a candidates importance, as evidenced by Eqs. 5 and 6, where the candidate model θ_{ca} is being defined.

We will use these three methods (BL, TFIDF, SEM, in that order) in combination with both Models 1 and 1B in our experimental evaluation (reported in Section 5).

3.3 Query Model

As to the query model, we consider two flavors. Our *baseline query model* (BL) consists of terms from the topic title only, and assigns the probability mass uniformly across these terms:

$$p(t|\theta_q) = p(t|q) = \frac{n(t, q)}{\sum_{t'} n(t', q)}. \quad (15)$$

As before, $n(t, q)$ is the frequency of term t in q.

The baseline query model has two potential issues. Not all query terms are equally important, hence, we may want to reweigh some of the original query terms. Also, $p(t|q)$ is extremely sparse, and, hence, we may want to add new terms (so that $p(t|\theta_q)$ amounts to query expansion). At this point, we consider yet another form of non-local evidence.

The TREC 2007 Enterprise track simulates a type of click-based system, where we have observed a given topic multiple times and where a small number of documents were often clicked. We refer to those documents as *example documents*. Balog et al. [7] propose an effective method of exploiting these example documents. Unlike previous work on relevance modeling [13] and blind relevance feedback mechanisms [21], here it is assumed that these expansion terms are sampled independently from the original query terms.

That is, we use a non-local approach to query expansion. The original (baseline) query model ($p(t|q)$) is combined with the expanded query model ($p(t|\hat{q})$) (EX) as follows.

$$p(t|\theta_q) = (1 - \mu) \cdot p(t|\hat{q}) + \mu \cdot p(t|q). \quad (16)$$

The expanded query is sampled from a set of example documents S. First, we estimate a "sampling distribution" $p(t|S)$ using example documents $d \in S$. Next, the top k terms with highest probability $p(t|S)$ are taken and used to formulate the expanded query $\hat{q}$:

$$p(t|\hat{q}) = \frac{p(t|S)}{\sum_{t' \in k} p(t'|s)}. \quad (17)$$

By summing over all example documents, we obtain $P(t|S)$. Formally, this can be expressed as

$$p(t|S) = \sum_{d \in S} p(t|d) \cdot p(d|s) \quad (18)$$

This resembles the way the candidate model is constructed in Eq. 3. We approximate $p(t|d)$ with the maximum-likelihood estimate and set $p(d|s)$ to be uniform (i.e., all example documents are equally important).

An example TREC topic, with the corresponding query models obtained using BL (Eq. 15) and EX (Eq. 16), is shown in Figure 1. We clearly see the reweighing and expansion effect of our new query model.

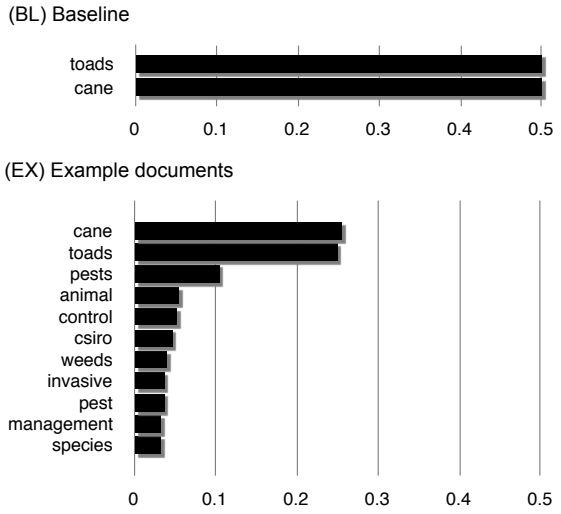

Figure 1: Query models generated for topic *CE-039: cane toads*

3.4 Candidate Priors

Our goal with introducing candidate priors is to demonstrate another form of incorporating non-local evidence into our modeling. Estimating this prior without training or manually encoding organizational knowledge is difficult (as was also found in [20]). We explored several approaches, including binning people by their document frequency or by the coherence of the set of documents in which they occur. While reasonably effective in distinguishing science communicators (and web masters and others whose names occur in many documents) from "proper" experts, we decided to use a simple pattern-based approach. We extracted a list names and positions within an organization from the *contact* blocks of documents (where this block existed). A large portion of these people are science communicators (SC) (often called *communication officer/manager/advisor* or *manager public affairs communication*).

We then set the candidate prior as follows:

$$p(ca) = \begin{cases} 1, & ca \notin SC, \\ 0, & ca \in SC. \end{cases} \quad (19)$$

This simply means that we identified science communicators and filtered them out from the list of names returned.

4. EXPERIMENTAL SETUP

4.1 Evaluation platform

To address our research questions (repeated below) we ran experiments using the CSIRO Enterprise Research Collection (CERC), a crawl of *.csiro.au (public) websites conducted in March 2007. The crawl contains 370,715 documents, with a total size 4.2 gigabytes [1].

In the 2007 edition of the TREC Enterprise track, CERC was used as the document collection [1]. CSIRO's science communicators played an important role in topic creation. They, the envisaged end-users of systems taking part in the TREC Enterprise track, read and create outward-facing web

pages of CSIRO to enhance the organization's public image and promote its expertise. A total of 50 topics were created by the science communicators; systems had to return "key contacts" for these topics, i.e., names that could be listed on the topic's overview page. These key contacts are considered as relevant experts, thus, used as the ground truth. It was not assessed whether there is evidence present in the collection to support the person's expertise.

4.1.1 Evaluation Measures

The measures we will use are (Mean) Average Precision (MAP), P5, P10 (precision at rank 5 and 10, respectively), and (Mean) Reciprocal Rank (MRR). MAP is appropriate since it provides a single measure of quality across recall levels. MAP is the main measure used for the expert finding task at the TREC Enterprise Track [11, 26, 2].

As to P5, P10, and MRR, we argue that recall (i.e., finding all experts given a topic or listing all expertise areas of a given person) may not always be of primary importance to our target users. Expertise retrieval can be seen as an application where achieving high accuracy, i.e., high precision in the top ranks is paramount. For this purpose P5, P10, MRR are appropriate measures [25].

4.2 Identifying Candidates

In the 2007 edition of the Expert Search task at TREC, candidates are identified by their primary e-mail addresses, which follow the `Firstname.Lastname@csiro.au` format. No canonical list of experts has been made available, therefore, e-mail addresses have to be extracted from the document collection, and then normalized to the primary format. This presents a number of challenges, including overcoming various spam protection measures, the use of alternative e-mail addresses, and of different abbreviations of names.

The list of candidates we use is taken from [3] and comprises 3,490 unique names in total. References of these people in documents were replaced by a unique identifier. See [3] for the description of the candidate extraction procedure.

4.3 Query Model Generation

We use the best performing query model from [7], EX-QM-EM, with $k = 30$ feedback terms. The original and the expanded query models are combined with equal weights: $\mu = 0.5$. All example documents were considered equally important ($p(d|S)$ is uniform).

4.4 Calculating Proximity

Note that our method for estimating the proximity-based model (Model 1B) allows for a weighted combination of various windows sizes (see Eq. 10). To remain focused, here we restricted ourselves to a single fixed window ($W = \{w\}$) and the size of this window was set to 125 (based best empirical results after performing a sweep on a set of possible window sizes from $20 \ldots 250$); see [3] for details.

4.5 Parameter Estimation

It is well-known that smoothing can have a significant impact on the overall performance of language modeling-based retrieval methods [27]. Our candidate models employ Bayes smoothing with a *Dirichlet prior* [18] to improve the estimated language models. Specifically, we set $\lambda = \frac{\beta}{\beta+|ca|}$, where $|ca|$ is the sum of the number of terms associated with a given candidate.

Based on an empirical investigation of smoothing values reported in [3] we set $\beta = 90,000$ for Model 1 and $\beta = 100$ for Model 1B.

5. EXPERIMENTAL EVALUATION

We repeat our research questions from the introduction and then present the results of the experiments performed to answer our questions.

5.1 Research Questions

We aim to find out to which extent rich query modeling with non-local evidence improves the effectiveness of expert finding: how do BL and EX compare across multiple experimental conditions. Second, we seek to determine how different ways of computing document-expert associations (with different types of global statistics) impacts expert finding: how do BOOL, TFIDF and SEM compare across multiple experimental conditions. And we determine to which extent priors on candidate experts (based on their global co-occurrence behavior) impact expert finding effectiveness.

5.2 Experimental Results

Table 2 lists the retrieval scores obtained for the various experimental conditions: the top half lists the scores based on Model 1, the bottom half lists the scores for Model 1B. Superscripts report on the outcome of significance tests (paired t-test, rows 2–6 vs. row 1, rows 8–12 vs. row 7, row 7 vs. row 1; $^{(1)} = .05$, $^{(2)} = .01$, $^{(3)} = .001$).

Model	$p(ca\|d)$	θ_q	MAP	P5	P10	MRR
1	BOOL	BL	.3801	.2000	.1340	.5571
	BOOL	EX	$.4518^{(1)}$	$.2360^{(2)}$	.1440	$.6481^{(1)}$
	TFIDF	BL	$.4478^{(2)}$	$.2520^{(2)}$	$.1580^{(2)}$	.6161
	TFIDF	EX	$.4957^{(2)}$	$.2800^{(3)}$	$.1640^{(2)}$	$.6861^{(1)}$
	SEM	BL	$.4541^{(2)}$	$.2440^{(1)}$	$.1580^{(3)}$	$.6252^{(1)}$
	SEM	EX	$.5044^{(3)}$	$.2720^{(3)}$	$.1640^{(2)}$	$.6866^{(2)}$
1B	BOOL	BL	$.4633^{(2)}$	$.2600^{(3)}$	$.1620^{(2)}$	.6236
	BOOL	EX	$.5178^{(1)}$	$.2840^{(1)}$	.1720	$.7009^{(1)}$
	TFIDF	BL	.4650	.2720	.1680	.6226
	TFIDF	EX	$.5380^{(1)}$	.2880	$.1800^{(1)}$	$.7064^{(1)}$
	SEM	BL	.4735	.2760	.1720	.6280
	SEM	EX	$.5465^{(3)}$	$.2880^{(1)}$	.1760	$.7119^{(1)}$

Table 2: Results overview. Document-candidate associations: (BOOL) Boolean, (TFIDF) Frequency-based using TF.IDF weighting, (SEM) Semantic relatedness; Query model: (BL) Baseline, (EX) Expanded (using example documents provided with the topic statement).

5.3 Query Models

How does rich (non-local) query modeling help expert finding? Moving from the baseline (BL) to more refined query formulations (EX) always improves and the improvement can be up to 19% in MAP, 18% in P5, 7% in P10, and 16% in MRR (even vs. odd rows of Table 2).

5.4 Document-Candidate Associations

How do (non-local) document-candidate associations help? Moving from local (BOOL) to more and more non-

local approaches (TFIDF and SEM) improves across the board and significantly, irrespective of the candidate and query models. On the other hand, the improvement gained by moving to non-local approaches is more substantial for Model 1 than for Model 1B.

5.5 Candidate Priors

Finally, we implement our candidate priors on top of the best performing configurations of Model 1 and Model 1B; see Table 3. Using priors result in significant improvements over these best performing configurations (MAP and MRR). Our scores reported in Table 3 outperform any previously published results that we are aware of.

Model	$p(ca)$	MAP	P5	P10	MRR
1	–	.5044	.2720	.1640	.6866
	SC	$.5506^{(2)}$	.2760	.1680	$.7344^{(2)}$
1B	–	.5465	.2880	.1760	.7119
	SC	$.5747^{(1)}$	$.3080^{(1)}$	.1780	$.7362^{(1)}$

Table 3: Results of adding candidate priors on top of the best performing configurations of Model 1 and Model 1B. Significance testing is done against these best forming configurations.

6. ANALYSIS

We start our analysis by contrasting the two extreme ends of the spectrum described in Table 2: a local approach M1-BOOL-BL (row 1 in Table 2) and a local approach mixed with non-local features, M1B-SEM-EX (row 12 in Table 2). Figure 2 shows a topic-level comparison. We find that in the majority of topics non-local features improve, and the improvement can be up to +.9655 Average Precision (AP) (topic *CE-015: life cycle assessment*). On the other hand, in a small number of cases it hurts performance—the rightmost bar corresponds to topic *CE-024: Double Helix Science Club* where AP drops by .4167. See Figure 2.

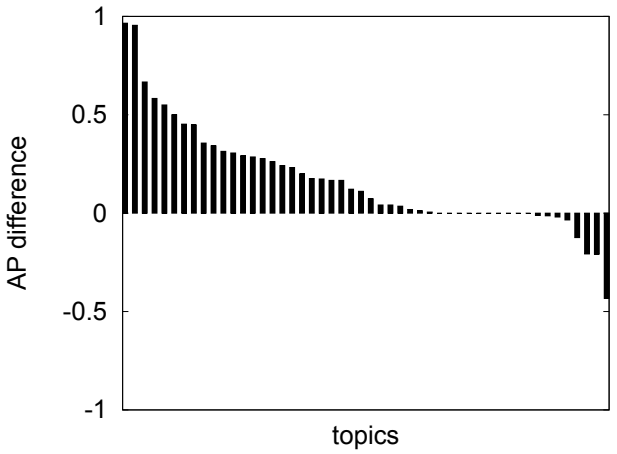

Figure 2: M1-BOOL-BL (baseline) vs. M1B-SEM-EX; row 1 vs row 12 of Table 2.

When we consider the move from Boolean to TFIDF-based document-candidate associations, we see that some topics are hurt, but on the whole more are helped by the move to TFIDF-based associations, independent of the query model being used (BL or EX); see Figure 3. The gains/losses (in numbers of topics) for the four pairwise comparisons shown in Figure 3 are: 30/10, 19/17, 31/8, 21/13, respectively.

Going from TFIDF to SEM we see that some topics are hurt, but more are helped, and by a bigger margin (Figure 4). The gains are more modest—both per topic and averaged—than the gains obtained by moving from BOOL to TFIDF (Figure 3). This is reflected in the gain/loss numbers: 17/15, 22/7, 12/16 (!), 15/9, respectively.

Next, we contrast runs with and without the expanded query model. Figure 5 shows the contrastive plots. On the whole, moving to richer query models has a positive effect, although some topics are hurt. Interestingly, we observe almost identical gain/loss patterns across Model 1 and Model 1B (top row vs. bottom row) and independent of the underlying association. The gain/loss numbers are (for plots (a)–(f)): 25/8, 24/12, 22/11, 20/12, 24/12, and 23/12, respectively.

Let us zoom in on the candidate models estimated using our document- and proximity-based models (Model 1 and 1B, respectively); Table 4 displays the terms associated with candidate *Manny Noakes* with the highest probability. Manny Noakes is leader of the research team that developed the *Total Wellbeing Diet*, published as a book (together with Dr Peter Clifton).[2] As we move from M1-BOOL to M1-SEM we can observe new terms emerging, such as *weight* and *nutrition*. Also, we can observe that several other associated terms move up in the ranking, e.g., *diet* and *health*. Switching from document-based to proximity-based models (i.e., from M1 to M1B) continues the progress in the direction of nutrition science, by adding terms like *protein* and *exercise*, while general terms, such as *industry* and *technology* have dropped out of the top 20. Finally, as we contrast M1B-BOOL and M1B-SEM, we observe slight refinements in the allocation of the probability mass; contrast, for example the probability of *nutrition* and *australia*.

Manny Noakes is an expert on topic *CE-013: human clinical trials* (according to the ground truth provided by CSIRO's science communicators). The two query models (BL, EX) for this topic are listed in Table 5. The ranking of Manny Noakes for this topic (using the different combinations of query model and candidate profile) is as follows:

Query mod.	M1-BOOL	M1-SEM	M1B-BOOL	M1B-SEM
BL	8	9	4	4
EX	6	4	3	3

As the query model gets richer, Manny Noakes' ranking improves, and, similarly, as the degree of non-locality improves. Given the query models and candidate profiles listed in Tables 5 and 4, we see why: the best performing models and profile are simply very similar.

Finally, when investigating the effect of candidate priors we find that these affect only a handful of topics, but the effect is always positive; see Figure 6.

According the literature (and our own previous publications), the document-based approach ("Model 2") was identified as a clearly preferred model as it is robust, is only slightly affected by smoothing and can be implemented efficiently on top of an existing document search engine [5, 8, 3]. However, when we contrast the numbers in Table 1 with the

[2] `http://www.csiro.au/people/Manny.Noakes.html`

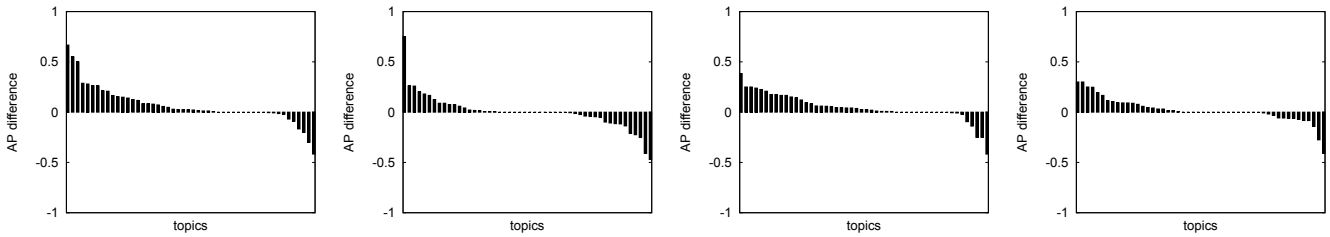

Figure 3: Moving from boolean (BOOL) to frequency-based associations (TFIDF), with the baseline query model (BL) or the expanded query model (EX). (a): Model 1 (BL); (b): Model 1B (BL); (c): Model 1 (EX); (d): Model 1B (EX). In terms of rows in Table 2: 1 vs. 3, 7 vs. 9, 2 vs. 4, and 8 vs. 10, respectively. Topics ordered by difference in AP.

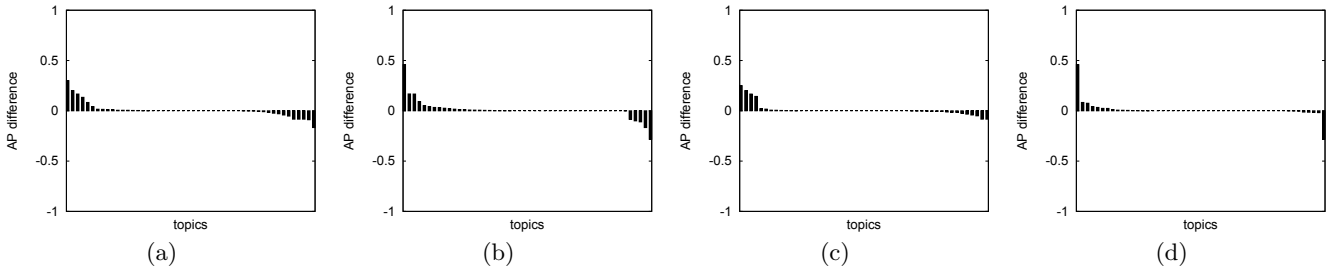

Figure 4: Moving from frequency-based (TFIDF) to semantic associations (SEM). (a): Model 1 (BL); (b) Model 1B (BL); (c): Model 1 (EX); (d): Model 1B (EX). In terms of rows in Table 2: 3 vs. 5, 9 vs. 11, 4 vs. 6, and 10 vs. 12, respectively. Topics ordered by difference in AP.

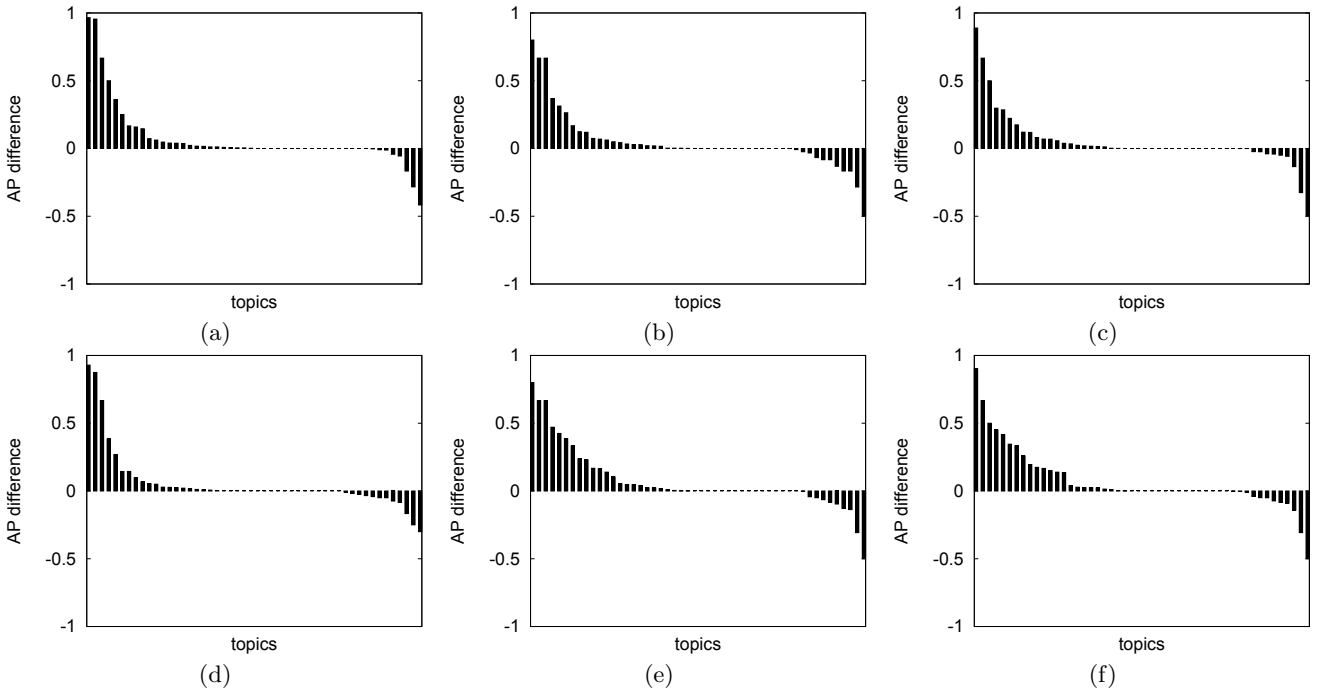

Figure 5: Moving from baseline query model (BL) to expanded query model (EX). Top row concerns Model 1, the bottom row Model 1B. (a): BOOL; (b) and: TFIDF; (c): SEM; (d): BOOL; (e): TFIDF; (f): SEM. In terms of rows in Table 2: 1 vs. 2, 3 vs. 4, 5 vs. 6, 7 vs. 8, 9 vs. 10 and 11 vs. 12, respectively. Topics ordered by difference in AP.

M1-BOOL		M1-SEM		M1B-BOOL		M1B-SEM	
t	$p(t\|\theta_{ca})$	t	$p(t\|\theta_{ca})$	t	$p(t\|\theta_{ca})$	t	$p(t\|\theta_{ca})$
csiro	.02217	csiro	.03608	csiro	.03048	csiro	.03125
food	.01144	diet	.01939	diet	.02524	diet	.02823
industry	.01011	wellbeing	.01302	wellbeing	.01685	wellbeing	.01839
diet	.00884	food	.01217	dr	.01630	total	.01600
research	.00808	total	.01078	total	.01450	dr	.01321
dr	.00764	health	.01005	research	.01002	health	.01132
australia	.00637	research	.00873	health	.00998	weight	.01028
wellbeing	.00608	energy	.00797	weight	.00871	book	.00959
science	.00554	industry	.00771	book	.00846	research	.00897
program	.00554	dr	.00742	australia	.00717	nutrition	.00833
health	.00542	australia	.00686	nutrition	.00704	high	.00666
total	.00526	science	.00623	food	.00619	food	.00661
new	.00475	weight	.00597	high	.00615	australia	.00653
energy	.00462	book	.00586	science	.00540	science	.00539
australian	.00425	information	.00584	protein	.00479	exercise	.00505
technology	.00387	technology	.00533	peterclifton	.00419	protein	.00490
information	.00371	high	.00521	loss	.00413	fat	.00489
2005	.00359	nutrition	.00514	new	.00407	peterclifton	.00466
development	.00337	flagship	.00479	exercise	.00406	information	.00431
management	.00327	program	.00448	team	.00402	adelaide	.00425

Table 4: Candidate models generated for *Manny Noakes*.

t	$p(t\|\theta_q)$	t	$p(t\|\theta_q)$
trials	.33333	trials	.19625
clinical	.33333	clinical	.18238
human	.33333	human	.17712
		csiro	.04328
		study	.03113
		adelaide	.02550
		participants	.02156
		health	.02020
		australia	.01988
		foods	.01871
		sheet	.01856
		information	.01856
		diet	.01669
		research	.01667
		food	.01590
		site	.01424
		particpants	.01293
		prospective	.01293
		woman	.01293
		young	.01293
		page	.01293
		questions	.01293
		based	.01293
		answers	.01293
		criteria	.01293
		contact	.01143
		nutrition	.00938
		functional	.00878
		participant	.00862
		obesity	.00860

Table 5: Query models generated for topic *CE-013: human clinical trials* (Left) BL, (Right) EX.

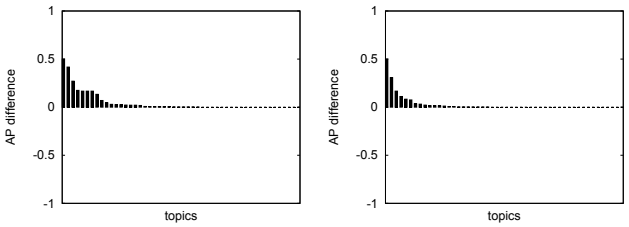

Figure 6: Topic level comparison when using the SC prior. (Left) M1-SEM-EX. (Right) M1B-SEM-EX. In terms of rows in Table 3: 1 vs. 2 and 3 vs. 4, respectively. Topics ordered by difference in AP.

best performing configurations we obtained in this section, we find that while Model 1 starts from a lower baseline, as additional non-local features are combined, it outperforms Model 2 and delivers state-of-the-art performance. We also added the non-local features discussed in this paper on top of Model 2, but this had only marginal effects [3].

We briefly summarize the pro and cons of Model 1. The cons include the need for maintenance of candidate models (as these have to be calculated offline, to be able to operate the retrieval system with an acceptable response time), and finding the optimal smoothing setting needs training material. Further, concerning Model 1B, calculating proximity could be done in more advanced ways (e.g., using proximity kernels as proposed in [20]). The pros include performance, and the fact that these models are "readable" for the user and can even be visualized as simply as tag-clouds.

7. CONCLUSIONS

We explored the use of non-local evidence for the task of expert finding. On top of existing document and proximity-

based language modeling approaches to the task, we considered three types of non-local evidence: obtained from query models, obtained from people-document associations, and as candidate priors. Starting from very competitive baselines we found that non-local evidence from query models helps improve expert finding effectiveness in all experimental conditions that we considered. Non-local aspects of document-candidate associations as modeled by the TFIDF approach improved over a Boolean baseline, while a semantics-based approach improved even more. On top of the best performing combinations of (non-local) query modeling and document-candidate associations, a final type of non-local evidence (candidate priors) leads to further improvements. Overall, our refined models outperform existing state-of-the-art approaches to expert finding.

Future work will concern ways of estimating within-document non-local evidence of expertise; many documents in the CSIRO test collection have additional (internal) structure, evidenced (among other things) by the presence of multiple text blocks—such blocks may be used to improve precision (just like proximity-based approaches), but at the same time evidence of associations between a candidate and a given topic may be scattered across multiple blocks: how can we identify text blocks that matter for candidate-topic associations?

8. ACKNOWLEDGEMENTS

We thank Wouter Weerkamp for helping us prepare this paper and our anonymous reviewers for their valuable feedback. This research was supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 017.001.190, 640.001.501, 640.002.501, 612.066.512, STE-07-012, 612.061.814, 612.061.815.

9. REFERENCES

[1] P. Bailey, N. Craswell, I. Soboroff, and A. de Vries. The CSIRO enterprise search test collection. *ACM SIGIR Forum*, 41, 2007.

[2] P. Bailey, N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC 2007 Enterprise Track. In *The Sixteenth Text REtrieval Conference Proc. (TREC 2007)*, 2008.

[3] K. Balog. *People Search in the Enterprise*. PhD thesis, University of Amsterdam, 2008.

[4] K. Balog and M. de Rijke. Associating people and documents. In *Proc. 30th European Conference on Information Retrieval*, pages 296–308, 2008.

[5] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proc. 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, 2006.

[6] K. Balog, T. Bogers, L. Azzopardi, A. van den Bosch, and M. de Rijke. Broad expertise retrieval in sparse data environments. In *Proc. 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 551–558, 2007.

[7] K. Balog, W. Weerkamp, and M. de Rijke. A few examples go a long way: Constructing query models from elaborate query formulations. In *Proc. 31th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 371–378, 2008.

[8] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expertise search. In *Information Processing & Management*, To Appear.

[9] 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.

[10] N. Craswell, D. Hawking, A. M. Vercoustre, and P. Wilkins. P@noptic expert: Searching for experts not just for documents. In *Ausweb*, 2001.

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

[12] H. Fang and C. Zhai. Probabilistic models for expert finding. In *Proc. 29th European Conference on Information Retrieval*, pages 418–430, 2007.

[13] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, 2001.

[14] C. Macdonald and I. Ounis. A belief network model for expert search. In *Proceedings of 1st conference on Theory of Information Retrieval (ICTIR)*, 2007.

[15] C. Macdonald and I. Ounis. Voting Techniques for Expert Search. *Knowledge and Information Systems*, 2007.

[16] C. Macdonald and I. Ounis. Expertise drift and query expansion in expert search. In *Proc. sixteenth ACM conference on Conference on information and knowledge management*, pages 341–350, 2007.

[17] C. Macdonald, D. Hannah, and I. Ounis. High quality expertise evidence for expert search. In *Proc. 30th European Conference on Information Retrieval*, 2008.

[18] D. J. C. Mackay and L. Peto. A hierarchical Dirichlet language model. *Nat. Language Engin.*, 1(3):1–19, 1994.

[19] D. Petkova and W. B. Croft. Hierarchical language models for expert finding in enterprise corpora. In *Proc. 18th IEEE International Conference on Tools With Artifical Intelligence (ICTAI'06)*, pages 599–608, 2006.

[20] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *Proc. sixteenth ACM conference on Conference on information and knowledge management*, pages 731–740, 2007.

[21] J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[22] H. Rode, P. Serdyukov, D. Hiemstra, and H. Zaragoza. Entity ranking on graphs: Studies on expert finding. Technical Report TR-CTIT-07-81, University of Twente, 2007.

[23] P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding'. In *Proc. 30th European Conference on Information Retrieval*, pages 309–320, 2008.

[24] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling relevance propagation for the expert search task. In *The Sixteenth Text REtrieval Conference Proceedings (TREC 2007)*, 2008.

[25] C. Shah and W. B. Croft. Evaluating high accuracy retrieval techniques. In *Proc. 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–9, 2004.

[26] I. Soboroff, A. de Vries, and N. Crawell. Overview of the TREC 2006 Enterprise Track. In *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, 2007.

[27] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342, 2001.

[28] J. Zhu, D. Song, S. Ruger, M. Eisenstadt, and E. Motta. The Open University at TREC 2006 Enterprise Track Expert Search Task. In *The Fourteenth Text REtrieval Conference Proceedings (TREC 2006)*, 2007.

[29] J. Zhu, D. Song, and S. Rüger. The Open University at TREC 2007 Enterprise Track. In *The Fifteenth Text REtrieval Conference Proceedings (TREC 2007)*, 2008.