Expertise Drift and Query Expansion in Expert Search

Craig Macdonald
University of Glasgow
Glasgow, Scotland, UK
craigm@dcs.gla.ac.uk

Iadh Ounis
University of Glasgow
Glasgow, Scotland, UK
ounis@dcs.gla.ac.uk

ABSTRACT

Pseudo-relevance feedback, or query expansion, has been shown to improve retrieval performance in the adhoc retrieval task. In such a scenario, a few top-ranked documents are assumed to be relevant, and these are then used to expand and refine the initial user query, such that it retrieves a higher quality ranking of documents. However, there has been little work in applying query expansion in the expert search task. In this setting, query expansion is applied by assuming a few top-ranked candidates have relevant expertise, and using these to expand the query. Nevertheless, retrieval is not improved as expected using such an approach. We show that the success of the application of query expansion is hindered by the presence of topic drift within the profiles of experts that the system considers. In this work, we demonstrate how topic drift occurs in the expert profiles, and moreover, we propose three measures to predict the amount of drift occurring in an expert's profile. Finally, we suggest and evaluate ways of enhancing query expansion in expert search using our new insights. Our results show that, once topic drift has been anticipated, query expansion can be successfully applied in a general manner in the expert search task.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.3.4 [Systems and software]: User profiles and alert services

General Terms: Performance, Experimentation

Keywords: Expert Finding, Expertise Modelling, Expert Search Information Retrieval, Query Expansion, Topic Drift

1. INTRODUCTION

In [16], Rocchio introduced the classical Information Retrieval (IR) concept of relevance feedback to improve a ranking of documents. An application of this is pseudo-relevance feedback (PRF), which has been used in adhoc search tasks to automatically improve the retrieval performance of document IR systems. The basic idea of PRF is to assume that a number of top-ranked documents are relevant, and learn from these documents to improve retrieval accuracy [19]. In Query Expansion[1] (QE), information from these top-ranked documents, known as the pseudo-relevant set, is used to expand the initial query and re-weight the query terms.

Unlike classical IR systems that generate rankings of documents, an *expert search* system aids a user in their "expertise need" by identifying people with relevant expertise to the topic of interest. Such a system can be useful in large Enterprise settings with vast amounts of digitised information, where people are a critical source of information because they can explain and provide arguments about why specific decisions were made [8]. Typically, an expert search system associates a set of documents to each candidate expert, known as profiles, to represent their expertise in the system. Candidates are then ranked in response to a query using the expertise evidence in their profiles.

In this paper, we aim to have a general application of query expansion (QE) to the expert search task, to enhance the retrieval accuracy of an expert search system. This aim is important, as while QE has been shown to be useful in adhoc document IR tasks [1, 15], the application of QE is not as useful for Web IR tasks, such as topic distillation and known-item finding [6]. In finding a general application of QE to the expert search task, we will show that it can indeed be successfully applied to increase the retrieval accuracy of an expert search system. Specifically, from an initial ranking of candidates with respect to a query, an application of QE in an expert search system would select several top-ranked candidate experts as the pseudo-relevant set, then expand the query using terms from their interests. Then re-running using the expanded query terms would generate a higher quality candidate ranking.

It is known that the effectiveness of QE in an adhoc document search system is affected by the quality of the initial top-ranked documents used for pseudo relevance feedback (known as the pseudo relevant set) [20]. However, we hypothesise that the presence of topic drift within candidate profiles can reduce the effectiveness of QE in the expert search task. What do we mean by this? Well a candidate expert can have several or many unrelated areas of expertise, which are reflected in the contents of their profile. Now we believe that when using the entire profile for query expansion for a query about a topic, these other unrelated expertise areas can wrongly influence the outcome of QE. Consider an IR example: W. B. Croft is generally considered an expert in language modelling, and an expert search system for IR should rank him highly in response to the query "language

[1]In this work, we use the terms pseudo-relevance feedback and query expansion interchangeably.

CIKM'07, November 6–8, 2007, Lisboa, Portugal.

modelling for IR". However Croft and other highly ranked candidates might share expertise in clustering. If QE is then applied, the expanded query terms might be more orientated towards clustering than language modelling, causing a topical drift in the new ranking of candidates.

This work aims to provide a framework for the general and successful application of QE in an expert search task. In particular, this work investigates the extent to which topic drift affects QE in expert search, and secondly, to investigate how to account for this expertise drift while applying QE in an expert search system.

The contributions of this paper are as follows: Firstly, we show how QE can be appropriately applied in an expert search task, when the pseudo-relevance objects are only candidate names. While this is useful for the successful automatic application of QE, this will also facilitate several other related tasks, for instance, the interactive application of QE in an expert search setting, or perhaps finding similar experts. Secondly, we propose and evaluate several measures for 'cohesiveness'. While these measures are evaluated in the context of the expert search task, it is of note that these may have applications in other areas of IR. For instance, cohesiveness measures may be applied on a ranking of documents to facilitate diversifying the top-ranked results in order to satisfy more types of users [17]. Lastly, we present several fine-grained QE approaches for expert search that reduce the occurrence of topic drift during QE, leading to a more effective QE that works on a list of candidates. Moreover, these approaches for QE are general and do not depend on the retrieval approach used for ranking the candidates.

The remainder of this paper is structured as follows: In Section 2, we introduce the expert search task and the Voting Model for expert search that we use in this work. Section 3 introduces how QE can be applied in this task, and presents the experimental setting and the baseline retrieval performances applied in this paper. Moreover, in Section 4, we investigate the extent to which topic drift is occurring during QE. In Section 5, we present three measures which we use to predict the amount of expertise drift within a candidate profile. Section 6 proposes and evaluates approaches for considering expertise drift when applying QE. We show that these successfully reduce topic drift and enhance the application of QE in the expert search task. In Section 7, we provide concluding remarks and ideas for future work.

2. EXPERT SEARCH

Modern expert search systems for Enterprise settings work by using documents to form the profile of textual evidence for each candidate expert [5]. The candidate's profile represents the expertise of the candidate expert in the expert search system. This documentary evidence can take many forms, such as intranet documents, documents, emails authored by the candidates, or web pages visited by the candidate (see [11] for an overview). In this work, the profile of a candidate is considered to be a set of documents associated with the candidate. These candidate profiles can then be used to rank candidates in response to a query.

Among the first models for expert search, is that proposed by Craswell et al [7], where all documents in each candidate's profile are combined into 'virtual documents', which are then directly ranked in response to a user query. However, because the contribution of each document in a profile is not measured individually, this approach is less effective than other subsequent approaches.

The advent of the expert search task in the recent TREC 2005 and 2006 Enterprise tracks has stimulated research interest in expert search [5, 18]. From this forum, there have been three main approaches for expert search: Balog et al. proposed the use of language models in expert search [3] based on two formal models. Their first model is based on Craswell et al's virtual document approach described above. For their second model, evidence from distinct documents in the candidate profiles are combined. Their experimental results showed that the second model improved over the simpler first model. Later, the probabilistic approach proposed by Cao et al. in [4] and the hierarchical language models proposed by Petkova & Croft [14] use a more fine-grained approach with windowing of documents around candidate name occurrences. However, in all three approaches, the relevance score of each candidate is determined utilising the relevance score of documents, as calculated using a language modelling approach.

In contrast, the Voting Model for expert search proposed by Macdonald & Ounis in [11] considers the problem of expert search as a voting process. The *ranking of documents*, with respect to the query Q, denoted by $R(Q)$, is assumed to provide inherent evidence about a possible ranking of candidates. The ranking of candidates can then be modelled as a voting process, from the retrieved documents in $R(Q)$ to the profiles of candidates: every time a document is retrieved and is associated with a candidate, then this is a vote for that candidate to have relevant expertise to Q. The ranking of the candidate profiles can then be determined by applying a voting technique that aggregates the votes of the documents. Eleven voting techniques for ranking experts were defined in [11]. Each of these voting techniques employ various sources of evidence derived from the ranking of documents, such as counting the number of documents associated with each candidate that are retrieved (number of votes), or the scores or ranks of the associated documents of each candidate (strength of votes).

In this work we choose to use the Voting Model for expert search proposed by Macdonald & Ounis , because it also takes the relevance of documents in each candidate's profile into account. Moreover, in contrast to [3], which can only use the language modelling approach, the Voting Model is general and flexible, and not limited to any document retrieval approach. In particular, we apply the expCombMNZ technique, which ranks candidates by considering the sum of the relevance scores of the documents associated with each candidate's profile, combined with the count of the number of documents from the profile that are ranked in the document ranking $R(Q)$. In expCombMNZ, the relevance score of a candidate C's expertise to a query Q is given by:

$$score\_cand_{expCombMNZ}(C,Q) = \|R(Q) \cap profile(C)\| \cdot \sum_{d \in R(Q) \cap profile(C)} exp(score(d,Q)) \quad (1)$$

where $profile(C)$ is the set of documents associated with candidate C, and $score(d,Q)$ is the relevance score of the document in the document ranking $R(Q)$. The number of documents from the profile of candidate C that are in the ranking $R(Q)$ is denoted by $\|R(Q) \cap profile(C)\|$, and $exp()$ is the exponential function. Note that this approach is general, as any retrieval model can be used to generate the initial ranking of document $R(Q)$. The $exp()$ function serves to bias the retrieved candidates towards those associated to higher-ranked documents.

Section 3 introduces the two ways that QE is applied in this work. The first of these is suitable for any expert search technique that uses candidate profiles as sets of documents, while the latter is only applicable using the Voting Model.

3. QE IN EXPERT SEARCH

The application of query expansion in adhoc search tasks is known to improve retrieval performance [1, 15]. To have a general application of QE to the expert search task, we desire to have a QE mechanism that works on the ranking of candidates (which we call Candidate Centric QE).

In [12], a candidate centric approach for QE was proposed that considers the entire profiles of the top-ranked candidates as the pseudo-relevant set. Note that this approach is not limited to the Voting Model, and can be applied to any expert search model that uses profiles to rank candidates. Moreover, an alternative application of QE, called Document Centric QE, was also proposed in the setting of the Voting Model, where the initial ranking $R(Q)$ is improved by the application of QE, before the voting technique generates the final ranking of candidates.

As mentioned above, to have a general application of QE for expert search, we need to consider the candidate ranking as the pseudo-relevant set. However, the experimental results from [12] show that the candidate centric approach to QE did not perform as well as the document centric approach. In this paper, we hypothesise that the failure exhibited by candidate centric QE is due to the occurrence of topic drift. Therefore, we aim to investigate and measure the topic drift problem in candidate centric QE, and then propose how candidate centric QE can be markedly improved by reducing its susceptibility to topic drift. If this topic drift is anticipated in candidate centric QE, and can enhance retrieval performance over the baselines, then we can conclude that QE can be successfully applied in the expert search task. In the remainder of this section, we introduce the core expert search experimental setting applied in this work, and the document centric and candidate centric baselines.

3.1 Experimental Setup

In this section, we define our experimental setup. Our experiments are carried out in the setting of the Expert Search tasks of the TREC Enterprise track, 2005 and 2006. The TREC W3C collection is indexed using Terrier [13], removing standard stopwords and applying the first two steps of Porter's stemming algorithm. Our initial experimental results have shown that applying only this weaker form of stemming results in increased high precision without degradation in mean average precision (MAP) for this task.

Next, we generate the profiles of documentary evidence of expertise for the candidates: for each candidate, documents which contain an exact match of the candidates full name are used as the candidate's profile. Using exact name matches, instead of say the candidates' last names only, ensures that only documents the candidates are definitely related to are associated with them, hence reducing the amount of mismatched evidence and ensuring good retrieval performance [3].

From the two TREC expert search tasks, we have a total of 99 topics with relevance assessments. For the TREC 2006 topics, where there are several topic fields, we only use the title field (ie short queries) - the TREC 2005 task only had one topic field where all terms formed the query. Documents in the initial ranking $R(Q)$ are ranked using the DLH13 document weighting model [10] from the Divergence from Randomness (DFR) framework [1]. We chose to experiment using DLH13 because it performs robustly on many collections and tasks (including expert search) without any need for parameter tuning [10, 11]. Indeed, DLH13 has no term frequency normalisation parameter that requires tuning, as this is assumed to be inherent to the model. Hence, by applying DLH13, we remove the presence of any term frequency normalisation parameter in our experiments.

In QE, terms found in the pseudo-relevant set are weighted, and the best of these are added to the initial query. In this work, we use the query expansion mechanism from the DFR framework [1]. In particular DFR deploys several term weighting models that measure the informativeness of each term in the pseudo relevant set. In our investigation into query expansion in expert search, we need to determine if the term weighting model employed has any effect on the conclusions concerning our two approaches for query expansion. DFR term weighting models measure the informativeness of a term by considering the divergence of the term occurrence in the pseudo-relevant set from a random distribution. One term weighting model, known as Bo1, is based on Bose-Einstein statistics and is similar to Rocchio [1]. The other is based on the Kullback Leibler (KL) divergence between the pseudo-relevant set sample and the collection. In Bo1, the informativeness $w(t)$ of a term t is given by:

$$w(t) = tf_x \cdot log_2 \frac{1 + P_n}{P_n} + log_2(1 + P_n) \quad (2)$$

where tf_x is the frequency of the term in the pseudo-relevant set, and P_n is given by $\frac{F}{N}$; F is the term frequency of the query term in the whole collection and N is the number of objects in the collection.

Alternatively, $w(t)$ can be calculated using a term weighting model based on Kullback Leibler divergence [1]:

$$w(t) = P_x \cdot log_2 \frac{P_x}{P_c} \quad (3)$$

where $P_x = \frac{tf_x}{l_x}$ and $P_c = \frac{F}{token_c}$. We denote by l_x, the size in tokens of the pseudo-relevant set, and $token_c$ denotes the total number of tokens in the collection. Using either Bo1 or KL to define $w(t)$, the top exp_term informative terms are identified from the top exp_item ranked items (these must exist in at least 2 items), and these are added to the query ($exp\_term \geq 1$, $exp\_item \geq 2$). Finally, for both the Bo1 and KL term weighting models, the query term frequency qtw of an expanded query term is given by [1] $qtw = qtw + \frac{w(t)}{w_{max}(t)}$, where $w_{max}(t)$ is the maximum $w(t)$ of the expanded query terms. $qtw = 0$ if the query term was not in the original query. We use the default setting for the QE parameters, ie $exp\_item = 3$ and $exp\_term = 10$, suggested by Amati in [1] after extensive experiments with several adhoc document test collections. While adjusting these parameters may enhance the retrieval performance of both document centric and candidate centric QE, we choose to leave these at their default settings, as initial experiments have shown that these do not alter the conclusions [12].

3.2 Results

Table 1 presents the retrieval performance achieved by the baseline expert search system, and by applying the document centric (DocQE) and candidate centric (CandQE) forms of QE, using both the Bo1 and KL term weighting models. The retrieval performance is reported on the

	TREC 2005		TREC 2006	
	MAP	P@10	MAP	P@10
Baseline				
	0.2037	0.3100	0.5502	0.6837
DocQE				
Bo1	**0.2185**	**0.3340***	**0.5606**	**0.6959**
KL	**0.2231***	**0.3400****	**0.5689***	**0.7020**
CandQE				
Bo1	0.1760	0.2500	0.4554	0.5939
KL	0.2031	0.3100	0.5600	0.6592

Table 1: Results for QE using the Bo1 and KL term weighting models. Results are shown for the baseline runs, with document centric query expansion (DocQE) and candidate centric query expansion (CandQE). The best results for each measure and term weighting model combination are emphasised. Statistically significant improvements at ($p \leq 0.05$) and ($p \leq 0.01$) over the corresponding baseline are denoted by * and **, respectively.

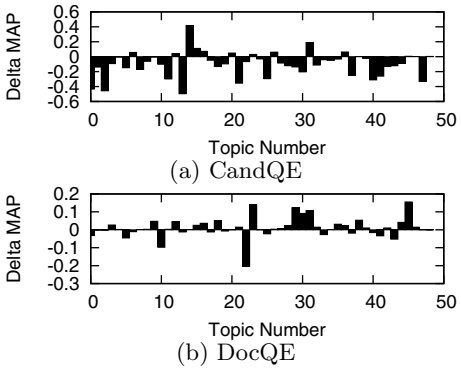
TREC 2005 and 2006 Enterprise track, expert search tasks. Statistically significant improvements from the baselines are shown using the Wilcoxon signed rank test.

Firstly, the performances from TREC 2005 and TREC 2006 are widely different. This follows the normal pattern: TREC 2005 was widely seen as a pilot task, where the candidate expertise relevance assessments were derived from an out-of-corpus ground truth - the membership of the W3C working groups. In contrast, for the TREC 2006 expert search task assessments were made for each pooled candidate, and hence the scale of the evaluation results is very different. The baseline voting technique, expCombMNZ, combined with the DLH13 document weighting model performs well above the median run for TREC 2005 (MAP 0.1402), and these results are similar to those of the 3rd top group participating that year. For TREC 2006, the median run of (MAP 0.3412), and these results are similar to those of the 2nd top group participating that year. Moreover, these settings are very competitive baselines on which to base our experiments.

With regards to the application of QE, at first inspection, it appears that document centric QE outperforms the candidate centric QE on both MAP and P@10, in all settings. In particular, it can be seen that applying document centric QE results in an increase over the baseline for both the TREC 2005 and TREC 2006 topics, on both MAP and P@10. These improvements are statistically significant ($p < 0.05$) in 4 out of 8 cases.

Finally, compared to the baselines, applying candidate centric QE almost always results in a degradation from the baselines, the exception being MAP for KL on the TREC 2006 topics, but this increase is not significant. In particular, Figure 1 shows the breakdown by topic of delta MAP for CandQE and DocQE (TREC 2006, Bo1). These show that while CandQE can enhance performance for some topics, it can seriously damage performance on many more topics. In contrast, DocQE improves many more topics. Moreover, across these topics, there is only a weak correlation (Spearman's $\rho = 0.267$, not statistically significant) between the topics that CandQE improves and those that DocQE improves.

The use of the Voting Model allows inference with respect to the document ranking $R(Q)$. In particular, it is intuitive



(a) CandQE

(b) DocQE

Figure 1: By topic breakdown of changes in MAP for CandQE and DocQE (with Bo1) on the TREC 2006 topics.

that a higher quality $R(Q)$ should improve the quality of the generated candidate ranking. These results infer that while DocQE improves $R(Q)$, there appears to be no benefit in applying QE directly in the expert search task, as concluded in [12]. However, this is counter-intuitive: the application of pseudo-relevance feedback has been shown to improve retrieval effectiveness in other adhoc tasks, so it seems there is a problem in the application of candidate centric QE that needs addressed. We hypothesise that the problem is that of topic drift occurring during candidate centric QE. In Section 4, we illustrate the extent to which topic drift is occurring. Moreover, in Section 5, we introduce and evaluate several 'cohesiveness measures' that attempt to predict when topic drift is occurring in a candidate profile. Our assumption which is evaluated in Section 6, is that if topic drift is accounted for in the candidate centric QE, then retrieval performance will be markedly improved.

4. CANDIDATE CENTRIC QE FAILURE ANALYSIS

We suggest that the less promising performance of candidate centric QE is due to 'topic drift'. A candidate profile contains many documents that represent the various interests of a candidate. As illustrated in Section 1 by the W. B. Croft example, when candidate centric QE is performed, the expanded query terms may describe other common, but not relevant, interests of the candidates in the pseudo-relevant set, causing more candidates with these incorrect interests to be retrieved erroneously. Topic drift is more likely to occur with candidate centric QE than with document centric QE as candidate profiles contain many documents, likely to be about several topics, while, comparatively, single documents are likely to remain related to one or two topics.

We develop two methods to measure the extent that topic drift is occurring during candidate centric QE. The first of these analyses the candidates that were used in the pseudo-relevant set. The second method investigates the quality of the expanded query terms.

By examining the relevance assessments for the expert search task, it is possible to observe that some candidates can have relevant expertise to multiple topics. Figure 2 shows the distribution of the number of topics candidates have relevant expertise in, for the TREC 2005 and TREC 2006 topics. Note that for TREC 2006, assessors were asked to judge for each topic the pooled candidates for relevance, using supporting documents to make those judgements. This was a substantially more complete judgement than for TREC

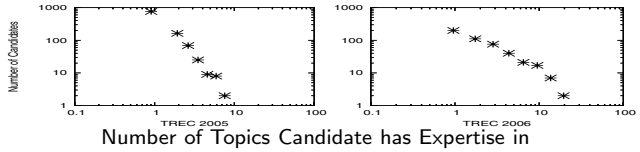


Figure 2: The distribution of the number of topics candidates have relevant expertise in, for the TREC 2005 and TREC 2006 relevance assessments.

2005, where relevance assessments were emulated using an out-of-corpus ground truth (W3C working ground membership). Hence, for the TREC 2005 set, we believe that the emulated judgements are incomplete from the viewpoint of the candidate - ie they do not reflect accurately the number of areas of expertise that many candidate have. Moreover, this can be observed in that there are a higher number of candidates that are only expert in one topic for the TREC 2005 set than the TREC 2006 set (see Figure 2).

To assess the extent that the candidates being used for relevance feedback in candidate centric QE had many areas of expertise, we count how many times they have been judged as relevant in the relevance assessments. The ideal scenario is that the candidates used in the pseudo-relevant set are not just expert in the current topic, but are not expert in any other topics, to prevent topic drift occurring during QE. For the reasons mentioned above regarding the number of expertise judgements in the TREC 2005 relevance assessments, we use only the TREC 2006 judgements for this analysis.

In fact over all 99 topics, for the candidate centric QE, the candidates used in the pseudo-relevant set were, in average, expert in 9.62 topics of interest. This is strikingly different from the average expertise of 1.27 topics for each candidate in the collection. This infers that the candidates used in pseudo-relevant set were expert in more topics than the current topic, and hence the QE mechanism was more likely to be affected by topic drift by identifying off-topic terms to expand the query with. Furthermore, by correlating the delta Average Precision in applying candidate centric QE over the no QE baseline with the average number of topics that pseudo-relevant candidates had interests in, we can indeed relate the problem of topic drift in the candidate profiles to poor QE performance. For instance, when using the Bo1 term weighting model on the 49 TREC 2006 queries, the correlation (Spearman's ρ) exhibited is $\rho = -0.357$, which is a statistically significant correlation. The negative correlation shows that when the candidates used in the pseudo-relevant set are expert in only few topics, QE is likely to do better, while if they are expert in many topics, it is likely that it is detrimental to apply QE to that query.

Our second measure examines the quality of the query terms added to the initial query by the QE approach. We compare the expanded query terms brought by the document centric and candidate centric QE approaches, by measuring the probability of an expansion term occurring in the relevance assessments. As the judgements for TREC 2006 were performed by identifying supporting documents for each relevant candidate [18], we use the set of supporting documents for all relevant candidates as our ground truth. From the results in Table 1, we expect that the expanded terms identified by candidate centric QE will not occur as much in the relevance judgements as those identified by document centric QE.

Formally, for a query Q which has expanded query terms

	Mean $P(Q_e\|Rel)$	
	Bo1	KL
CandQE	$2.55 * 10^{-3}$	$3.91 * 10^{-3}$
DocQE	$3.54 * 10^{-3}$	$4.32 * 10^{-3}$

Table 2: For the TREC 2006 setting, the mean probability of an expanded query Q_e being generated by the relevant supporting documents (Mean $P(Q_e|Rel)$).

Q_e, the probability of the expanded query terms occurring in the set of relevance assessments for query Q, Rel, is:

$$P(Q_e|Rel) = \frac{1}{exp\_term} \cdot \sum_{t \in (Q_e)} qtw \cdot \frac{tf_{Rel}}{token_{Rel}} \quad (4)$$

where tf_{Rel} is the term frequency of term t in the set of document Rel, and $token_{Rel}$ is the number of tokens in the set Rel. exp_term is the number of expanded query terms. qtw is the weight given to the expanded query term t in the refined query. It is used to prevent query terms that were given little weight in the expanded query biasing the measure.

Table 2 presents the Mean $P(Q_e|Rel)$ for each QE setting on the TREC 2006 topics. From this we can see that the likelihood of the expanded terms being in the relevant document is lower for both candidate centric QE settings. This demonstrates that indeed the query terms being identified in candidate centric QE are less useful than those identified by document centric QE. Because of the nature of the applied term weighting models, we reject the idea that they are identifying noise as informative terms, and instead hypothesise that that topic drift is indeed occurring in the candidate centric QE, compared to the document centric QE.

In the following section, we investigate how we can automatically predict the extent to which a candidate profile is about one central area of expertise. Following Amitay et al [2], who measured the 'cohesiveness' of a ranking of documents, we denote a candidate profile in which the expert has one sole interest as cohesive. In the following section, we present three ways of measuring cohesiveness, of which two are from the vector-space and language modelling frameworks. Our aim is that if we can show that un-cohesive candidate profiles can be identified, then we can possibly take this into account for an enhanced QE approach.

5. MEASURING COHESIVENESS

In the previous experiments, we hypothesise that the expertise drift within a candidate profile are responsible for the poor performance of the candidate centric QE. To this end, we investigate how cohesive a candidate's profile is: we measure the extent to which a candidate's expertise profile is around a central topic. For this we use three measures: firstly, simply counting the number of documents associated with each candidate ($\|profile(C)\|$), secondly the Cosine measure, and lastly Kullback-Leibler (KL) divergence [9].

For the first of these measures, $\|profile(C)\|$, our intuition is simply that the more expertise evidence found for a candidate, the more likely it is that the candidate's expertise varies across more than one topic. Moreover, as any expert search system must have knowledge of the documents in each candidate's profile, this measure is simple to calculate.

Our second and third cohesiveness measures are based on the intuition that the more the language model of a candidate's profile differs from its constituent documents, the less

cohesive the profile is. We use Cosine and KL divergence to measure this. The cohesiveness of a candidate profile can be measured using the Cosine measure from the vector-space framework as follows:

$$Cohesiveness_{Cos}(C) = \frac{1}{\|profile(C)\|} \cdot \sum_{d \in profile(C)} \frac{\sum_{t \in profile(C)} tf_d \cdot tf_C}{\sqrt{\sum_{t \in d}(tf_d)^2}\sqrt{\sum_{t \in profile(C)}(tf_C)^2}} \quad (5)$$

where tf_d is the term frequency of term t in document d, and tf_C is the total term frequency of term t in all documents in $profile(C)$ (denoted $t \in profile(C)$). $Cohesiveness_{Cos}$ measures the mean divergence between every document in the profile and the profile itself. Note that $Cohesiveness_{Cos}$ is bounded between 0 and 1, where 1 means that the documents represent the profile completely.

Alternatively, we measure the cohesiveness of a candidate profile by measuring the mean KL divergence between the language model of every documents in the profile and the language model model of the profile itself. Formally, the KL divergence between two probability distributions $\ominus_1$,$\ominus_2$ is:

$$KL(\ominus_1 \| \ominus_2) = \sum_t p(t|\ominus_1) log \frac{p(t|\ominus_1)}{p(t|\ominus_2)} \quad (6)$$

We use maximum likelihood to estimate the probability of a term t occurring in the document model $\ominus_d$, and the probability of a term in the profile model $\ominus_C$. To measure the cohesiveness of a candidate profile, we use the mean KL divergence between the language model of every document in the profile and the language model of the profile itself:

$$Cohesiveness_{KL}(C) = \sum_{d \in profile(C)} \frac{KL(\ominus_d \| \ominus_C)}{\|profile(C)\|} \quad (7)$$

Note that $\forall C, Cohesiveness_{KL}(C) \geq 0$, and the larger the value, the less cohesive the profile of candidate C is. We now evaluate the three defined measures of cohesiveness.

5.1 Evaluation

To evaluate our measures of cohesiveness, we use the relevance assessments of the TREC 2006 expert search task, as described in Section 4, as the ground truth to evaluate how effective we are at measuring the cohesiveness of candidates. Our hypothesis is that candidates with less cohesive candidate profiles (i.e. more expertise drift) will be expert in more topics, according to the relevance assessments, and will be more likely to cause topic drift in candidate centric QE. To perform the evaluation, we rank all the candidates which are expert in one or more topics, and correlate these with the cohesive measures defined above.

Table 3 shows the Spearman's rank correlation (ρ) between the cohesiveness measures and the ground truth from the TREC 2006 judgements. From the results, we can see that there are moderately strong correlations between all three cohesiveness measures and the ground truth, the highest of which is exhibited by $\|profile(C)\|$. Note that the correlation for $Cohesiveness_{Cos}$ is negative because this measure gives lowest values for the most cohesive profiles.

Furthermore, there are several possible reasons that an even higher correlation is not observed: Firstly, with only 49 topics from TREC 2006, it is entirely possible that some candidates expertise areas were not covered by the topics. This could mean that candidates predicted to have many areas of expertise are ranked low in the ground truth because the topics did not cover many of their expertise areas. Secondly, expertise assessment at TREC is performed by pooling the suggested candidates by submitted retrieval systems. This infers that not all possible candidates will have been judged for each topic, meaning that there may exist relevant candidates not judged. Thirdly, before an assessor can judge a candidate expert as having relevant expertise to the topic, they must have seen at least one supporting document. Supporting documents for each candidate are provided by systems, and are pooled for each candidate. A candidate who has relevant expertise in 'real life' may not be marked as relevant as a supporting document was not present in the collection, or not pooled and judged.

Cohesiveness Measure	ρ Correlation
$\|profile(C)\|$	0.585
$Cohesiveness_{Cos}(C)$	-0.517
$Cohesiveness_{KL}(C)$	0.566

Table 3: Correlations between various predictors of cohesiveness and the ground truth based on the TREC 2006 expertise relevance assessments.

Despite the caveats in this evaluation described above, the correlations exhibited in Table 3 demonstrate that these measures are sufficiently accurate with respect to the ground truth, and moreover, they are are equally comparable.

Other methods of measuring cohesiveness exist: For instance, in TREC 2003, Amitay et al. [2] filtered a ranking of documents for cohesive documents using the combination of IDF and Entropy. Alternatively, taking the mean divergence between every pair of documents in a candidate profile would have required the use of symmetric divergence operators, e.g. J-Divergence [9], however as some candidate profiles are extremely large (around 5000 documents), the time taken to compute such measures for all candidates would have been unfeasible. Indeed, some preliminary experiments suggest that 587,436,281 document-document comparisons would be required to measure the cohesiveness of all candidates in the profile set applied in this work[2].

Similarly, and analogous to [17], cohesiveness could be measured by clustering candidate profiles: the number of distinct clusters in a profile gives an indication of the number of topics the candidate showed expertise in. However, the simple measures proposed above give good correlations to our ground truth, and the most effective, $\|profile(C)\|$, is extremely cheap to compute, as an expert search system will already know the associations between documents and candidates. Now that we have reasonably good predictors of cohesiveness, we show in Section 6 how candidate centric QE can be improved to account for topic drift.

6. IMPROVING QE FOR EXPERT SEARCH

In the previous section, we proposed three measures which can predict how many topics a candidate has relevant expertise in. Moreover, when a candidate has many areas of expertise represented in their candidate profile, then this may be responsible for the occurrence of topic drift during candidate centric QE: if any additional non-relevant topic areas

[2]However, while 11% of these comparisons are duplicates and could be skipped, the time taken to compare this many document pairs would still be unfeasible for any real world applications or experimental settings.

were shared in the profiles of any candidates in the pseudo-relevant set, then terms from these topics areas might be added to the expanded query, causing candidates who only have expertise in the non-relevant topic areas to be retrieved.

In this section, we propose three approaches that enhance candidate centric QE, based on hypotheses concerning how topic drift can be reduced. The approaches are designed to reduce the topic drift that has been identified and discussed in this paper, and could be applied using other expert search techniques rather than the Voting Model.

Hypothesis 1: Query expansion can be enhanced by not considering candidates with non-cohesive profiles during pseudo-relevance feedback.

Hypothesis 2: Query expansion can be enhanced by only considering the on-topic parts of candidate profiles.

Lastly we combine Hypotheses 1 & 2 to form a third:

Hypothesis 3: Query expansion can be enhanced by only considering the on-topic parts of the non-cohesive profiles.

The remainder of this section defines three approaches for query expansion in the expert search task based on the three hypotheses respectively. The first of these approaches, *Selective Candidate Centric QE*, makes use of a measure of cohesiveness, such as those defined in Section 5 above, to prevent non-cohesive candidate profiles being considered for the pseudo-relevant set. We assume that by removing non-cohesive candidate profiles from the pseudo-relevant set, only candidates with relevant expertise *mostly about* the topic will remain. Expanding the query using this refined pseudo-relevant set would exhibit less topic drift than the candidate centric QE defined in Section 3. However, a possible disadvantage is that this approach is too harsh, and removes useful candidates from the pseudo-relevant set.

In contrast, the second approach (based on Hypothesis 2), *Candidate Topic Centric QE*, does not make use of the cohesiveness measures, but instead considers only the subset of documents in the candidate profiles which are about the initial user topic for inclusion in the pseudo-relevant set, similar to [19]. We can use the relevance score of the document to the query as an indicator for the topicality of each document in a candidate profile. By only considering the highest scored documents in the pseudo-relevant set of candidate profiles, the expanded query terms are more likely to be about the topic of interest. However, it is possible that the removed portion of the profile was a good source of expanded query terms.

Lastly, in the third approach, which we call *Selective Candidate Topic Centric QE* (Hypothesis 3), for the pseudo-relevant set we consider all the documents of the profiles of cohesive candidates, while for non-cohesive candidates, only documents from the profiles which are on-topics are considered. Similar to Selective Candidate Centric QE, we use a cohesiveness measure to predict the cohesiveness of the candidate profiles of the pseudo relevant set.

To show that we have successfully taken into account the topic drift, we compare to the CandQE results in Table 1. Moreover, to assess whether QE is actually useful in expert search, we compare also to the baseline (no QE) and to the stronger DocQE results from Table 1.

6.1 Selective Candidate Centric QE

In Hypothesis 1, we desire to reduce the amount of topic drift occurring during query expansion, which occurs because the candidate profiles used as the pseudo-relevant set are not cohesive. In this approach, which we denoted Selective Candidate Centric QE, we take into account a cohesiveness measure, such as one of these we defined in Section 5, to predict candidates that do not have a cohesive profile and hence should not be considered during QE.

We use the $\|profile(C)\|$ cohesiveness measure because this shows the highest correlation with our ground truth. For this approach, we set a threshold $sel\_profile\_docs$. When a candidate's profile contains more documents than the threshold $sel\_profile\_docs$, the candidate will not be considered for pseudo-relevance feedback.

Table 4 shows the results when applying Selective Candidate Centric QE while varying the $sel\_profile\_docs$ threshold. From the results, we can see that this approach for QE produces marked increases in both MAP and P@10 over the candidate centric QE baselines, some of these increases being statistically significant. Compared to the document centric baseline, improvements are exhibited on the TREC 2005 topics only (significant only in one case). With relation to the threshold $sel\_profile\_docs$, a value around 200 to 500 document appears to be a good setting for this collection. In particular, at a threshold of 500 on the TREC 2006 queries, the average number of TREC 2006 topics that the candidates in the pseudo-relevant set are expert in is 3.6. This is a marked contrast from the 9.62 topics observed in Section 4 for CandQE, and shows that the profiles used in this approach are much more cohesive. Contrasting the performance of the approach on the TREC 2005 and 2006 tasks, we see that more statistically significant increases are exhibited for the 2005 task, while the easier 2006 task shows a lesser benefit in applying this approach. This mirrors the results of DocQE and CandQE in Table 1. Finally, the underlined values in Table 4 show when selective candidate centric QE improves over all other settings for that task and term weighting model (baseline, DocQE, CandQE). We can see that the proposed simple approach is comparable to document centric QE for the TREC 2006 task, and outperforms it for certain threshold values on the TREC 2005 data.

6.2 Candidate Topic Centric QE

In Hypothesis 2, we desire to reduce the occurrence of topic drift when applying QE, by reducing the amount of irrelevant information in the candidate profiles considered for pseudo-relevance feedback. This is similar to how the language modelling [3] and voting approaches for expert search [11] improve over the virtual document approach of [7] - instead of focusing on the entire candidate profiles, the emphasis is placed on the on-topic documents within each candidate profile. When QE is being applied, it is unlikely that documents in the profiles which were not at least on-topic will bring any terms related to the user's topic of interest. Hence, they should not be considered for the pseudo-relevant set. In this case, the pseudo-relevant set for QE becomes the set of documents that are associated with the first exp_item ranked candidates, but are predicted to be relevant to the topic. We call this approach Candidate Topic Centric QE.

In this approach, we remove off-topic material from the pseudo-relevant set of candidate profiles before QE takes place. Detecting whether a document is on-topic is measured simply by using the relevance score of the document to the query, $score(d, Q)$. However, as most document weighting models do not compute bounded retrieval scores, we simply select the exp_cand_doc top scored documents from each of the candidate profiles in the pseudo-relevant set. The special value `ALL` designates when all documents with

	TREC 2005		TREC 2006	
sel_profile_docs	MAP	P@10	MAP	P@10
No QE	0.2037	0.3100	0.5502	0.6837
Bo1				
DocQE	0.2185	0.3340	0.5606	0.6959
CandQE	0.1760	0.2500	0.4554	0.5939
100	0.2222*†	0.3520*†	0.5041	0.5980
200	0.2387**†	**0.3780****†	**0.5240****	**0.6163**
500	**0.2477****†‡	**0.3780****†	0.5077	0.6102
1000	0.2191*†	0.3560**†	0.4857	0.5510
2000	0.2044†	0.3200*†	0.4894	0.6000
KL				
DocQE	0.2231	0.3400	0.5689	0.7020
CandQE	0.2031	0.3100	0.5600	0.6592
100	0.2228†	0.3500†	0.5415	0.6429
200	**0.2448***†	**0.3640**†	0.5549†	0.6592
500	0.2393**†	0.3580†	0.5540†	0.6571
1000	0.2136†	0.3220†	**0.5616**†	**0.6592**
2000	0.2015	0.3120†	0.5563†	0.6531

Table 4: Selective Candidate Centric QE: Candidates with $\|profile(C)\| \geq sel\_profile\_docs$ are not considered for pseudo-relevance feedback. The corresponding no QE, DocQE and CandQE baselines from Table 1 are included. The best results for each measure and term weighting model combination are emphasised. Statistically significant improvements at ($p \leq 0.05$) and ($p \leq 0.01$) over the corresponding CandQE run are denoted by * and **, respectively; † denotes when the measure is better than the CandQE baseline, while underlined values are better than the both best corresponding DocQE and no QE baseline results. Significant improvements over the DocQE baseline are denoted ‡.

$score(d, Q) > 0$ in the candidate profile are considered. Note also, that this approach is not specific to the Voting Model, as any expert search approach would be able to compute a relevance score for each document in a candidate's profile.

Table 5 presents the experimental results when applying candidate topic centric QE. We vary *exp_cand_doc* across a selection of values while the *exp_item* and *exp_term* QE parameters remain unchanged as in Section 2. Firstly, it is apparent that this approach for QE generates substantial increases on both TREC expert search tasks, for all measures and QE term weighting models. There are statistically significant increases in each setting for certain values of the threshold, and marked increases over the document-centric QE baseline, except for the KL term weighting model on the TREC 2006 task. In particular, the value 500 is very close to the `ALL` setting, and produces no difference in performance. The setting range 5-20 documents produce the best results on both tasks. Moreover, performance is enhanced more for MAP than P@10.

6.3 Selective Candidate Topic Centric QE

In this approach, similar to Selective Candidate Centric QE, a selective technique based on a cohesiveness measure. The aim here is to identify the uncohesive candidates in the pseudo relevant set, and reduce the topic drift that they induce, by applying Candidate Topic Centric QE only for those candidates. For the candidates with cohesive profiles, this filtering of the profile is unnecessary and is not applied.

Table 6 presents the experimental results when applying selective candidate topic centric QE across a range of settings of the *sel_profile_docs* threshold of the cohesiveness measure. In these experiments, we leave $exp\_cand\_doc = 10$, as this value gave good performance with the Candidate Topic Centric QE approach for both the TREC 2005 and 2006 tasks. Examining Table 6, we draw the following observations: firstly, this approach is also successful at improving over the CandQE baseline. Moreover, most settings on both TREC tasks can outperform the DocQE approach defined earlier, for both MAP and P@10 measures. With respect to parameter *sel_profile_docs*, the approach seems to be stable, with this having only some impact on retrieval performance, however the values 200 & 500 exhibit the best retrieval performance. Finally, the performance of the Selective Candidate Topic Centric QE approach would be improved by an appropriate setting of both parameters *sel_profile_docs* and *exp_cand_doc*. More generally, it would be useful to understand the connection between *exp_item*, *exp_cand_doc* & *exp_term* in a similar manner to the parameter scans presented in [12].

6.4 Discussion & Analysis

The approaches for query expansion described are general models for applying QE in expert search. Any of them could easily be applied using other term weighting models, or from candidate rankings generated using other expert search approaches. Comparing to the no QE baseline system defined in Section 3, the proposed QE approaches markedly outperform the baseline[3], suggesting that it is helpful to appropriately apply QE in expert search. In particular, MAP is generally improved, by applying the proposed approach for QE, however P@10 is less improved. This suggests that applying QE increases the recall of relevant candidates at lower ranks more than perfecting the top-ranked candidates. It is also of note that the proposed approaches are at least comparable to document centric QE, and in some cases exhibit a marked increase in performance.

Comparing the two first approaches, we can see that the approach based on Hypothesis 1 could be too harsh as it may remove the only useful expertise evidence for relevance feed-

[3]Denoted by underline in Tables 4, 5 & 6

	TREC 2005		TREC 2006	
exp_cand_doc	MAP	P@10	MAP	P@10
No QE	0.2037	0.3100	0.5502	0.6837
Bo1				
DocQE	0.2185	0.3340	0.5606	0.6959
CandQE	0.1760	0.2500	0.4554	0.5939
ALL	0.2039*†	0.2940†	0.5445**†	0.6506**†
5	**0.2240**** †	0.3400†	0.5381**†	0.6531**†
10	0.2174**†	**0.3260**†	0.5522**†	**0.6265**** †
20	0.2194**†	0.3160†	**0.5567**** †	0.6551**†
50	0.2142*†	0.3160†	0.5355**†	0.6265**†
100	0.2062†	0.3020†	0.5436**†	0.6347**†
200	0.2034†	0.2940†	0.5452**†	0.6367**†
500	0.2039†	0.2940†	0.5445**†	0.6306**†
KL				
DocQE	0.2231	0.3400	0.5689	0.7020
CandQE	0.2031	0.3100	0.5600	0.6592
ALL	0.2235†	0.3340†	0.5672†	0.6776†
5	**0.2314**** †	**0.3520**†	0.5582†	0.6831†
10	0.2255*†	0.3480†	0.5702†	0.6776†
20	0.2215†	0.3380†	**0.5749**†	0.6857†
50	0.2246*†	0.3380†	0.5693†	**0.6857**†
100	0.2228†	0.3260†	0.5675†	0.6673†
200	0.2231†	0.3240†	0.5677†	0.6755†
500	0.2235†	0.3340†	0.5672†	0.6776†

Table 5: Candidate Topic Centric QE: Only the top *exp_cand_doc* highest ranked documents in each candidate's profile are considered for pseudo-relevance feedback. Notation as Table 4.

	TREC 2005		TREC 2006	
sel_profile_docs	MAP	P@10	MAP	P@10
No QE	0.2037	0.3100	0.5502	0.6837
Bo1				
DocQE	0.2185	0.3340	0.5606	0.6959
CandQE	0.1760	0.2500	0.4554	0.5939
100	0.2135**†	0.3320**†	0.5608**†	0.6857**†
200	0.2159**†	0.3380**†	0.5614**†	0.6837**†
500	0.2299**†	0.3500**†	**0.5688**** †	**0.6918**** †
1000	**0.2356**** †	**0.3600**** †	0.5644**†	0.6837**†
2000	0.2125**†	0.3200**†	0.5392**†	0.6510**†
KL				
DocQE	0.2231	0.3340	0.5689	0.7020
CandQE	0.2031	0.2500	0.5600	0.6592
100	0.2257**†	0.3340†	0.5681†	**0.6959**†
200	**0.2271**** †	0.3380†	0.5739†	0.6918†
500	0.2245*†	**0.3400**†	**0.5783**†	0.6918†
1000	0.2235*†	0.3340†	0.5696†	0.6776†
2000	0.2205†	0.3380†	0.5721†	0.6755†

Table 6: Selective Candidate Topic Centric QE: For candidates with $\|profile(C)\| < sel\_profile\_docs$, the pseudo relevance set includes all documents from their profile, while for candidates with un-cohesive profiles (ie $\|profile(C)\| \geq sel\_profile\_docs$), only the top *exp_cand_doc* highest ranked documents in each candidate's profile are considered for pseudo-relevance feedback. In this table, $exp\_cand\_doc = 10$. Notation as Table 4.

back. The approach based on Hypothesis 2 relies heavily on the quality of the document relevance scores. According to the results in Tables 4 & 5, there is no clear winner over both years of the TREC tasks: for TREC 2005, both approaches are equivalent; while for the TREC 2006 task, Candidate Topic Centric QE performs best overall.

The Selective Candidate Topic Centric QE approach presented in Table 6 is a stable approach that consistently outperforms the CandQE baseline (except for P@10 using Bo1 on the TREC 2006 queries)[4], and outperforms the other approaches. Moreover, we wish to check that the approach does not favour longer or shorter candidates than the original baseline expert search system. To this end, we devise a measure that sees how prolific the candidates in the ranking are:

[4]Denoted by † in Tables 4, 5 & 6

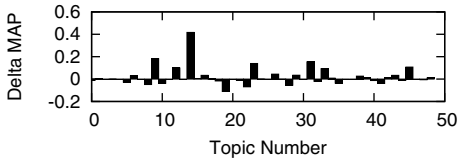


Figure 3: By topic breakdown of changes in MAP for Selective Candidate Topic Centric QE (with Bo1 and $sel\_profile\_docs = 500$) on the TREC 2006 topics.

$$Prolificness\_Ranking(Q) = \sum_{C \in Q} \frac{Topics(C)}{rank(C,Q)} \qquad (8)$$

where $C \in Q$ is all the candidate retrieved for query Q, $Topics(C)$ is the number of topics from the TREC 2006 task that the candidate C was expert in, and $rank(C,Q)$ is the rank of candidate C in the ranking for query Q. In particular, for the TREC 2006 task, using the setting $sel\_profile\_docs = 500$ for the Bo1 term weighting model on the Selective Candidate Topic Centric QE approach, the mean $Prolificness\_Ranking$ across all queries is 37.3, which is very similar with the baselines: no QE (37.5); CandQE (37.2); and DocQE (36.6). Hence, we conclude that no additional bias towards candidates with a few or many interests has been introduced by this approach.

To investigate how similar it is to the two query expansion baselines (CandQE and DocQE), we calculate the delta MAP from each approach to the baseline, and then correlate. In particular, for the TREC 2006 task, using the setting $sel\_profile\_docs = 500$ for the Bo1 term weighting model on the Selective Candidate Topic Centric QE approach, we find that the delta MAP is more closely correlated with the CandQE baseline than the DocQE baseline ($\rho = 0.518$ vs $\rho = 0.305$). This is surprising given that the new approach outperforms the CandQE baseline by +24.9%. By comparing Figure 3, which shows a histogram of the delta MAP given by the new approach, to Figure 1, it can be seen that the new approach improves on CandQE by eliminating many of the negative queries of CandQE, whilst keeping the better queries.

7. CONCLUSIONS & FUTURE WORK

In this paper, we showed that dealing with the topic drift problem is necessary for a succesfull application of query expansion in expert search. In particular, when topic drift is appropriately dealt with, applying QE can improve on a no QE baseline, and on a simple QE applied on the retrieved documents (DocQE). Moreover, with appropriate settings of the parameters in the proposed approaches, further enhancement of retrieval performance is likely. Lastly, the proposed approaches can be easily implemented on top of an existing document search engine, without the need for any additional index structures.

The cohesiveness measures proposed in this work have applications other than in the expert search task. For instance, in a normal search engine, it may be desirable to produce a diverse ranking of documents for ambiguous queries, to satisfy more possible distinct user needs [17].

Our future research directions from this work are two-fold: Firstly, it is clear that the problem of topic drift does occur, particularly within the expert search task. Further measures that can show when and how topic drift is occurring during QE would be beneficial. Secondly, the successfull application of QE to expert search introduces other potential applications, such as finding similar experts, creating a diverse ranking of candidates for ambiguous queries, and even the automatic creation of a 'roadmap of expertise' in an organisation.

8. REFERENCES

[1] G. Amati. *Probabilistic Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, Univ. of Glasgow, 2003.

[2] E. Amitay, D. Carmel, A. Darlow, M. Herscovici, R. Lempel, A. Soffer, R. Kraft, and J. Y. Zien. Juru at TREC 2003 - Topic Distillation using Query-Sensitive Tuning and Cohesiveness Filtering. In *TREC 2003*, pages 276–282.

[3] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of SIGIR 2006*, pages 43–50.

[4] Y. Cao, H. Li, J. Liu, and S. Bao. Research on Expert Search at Enterprise Track of TREC 2005. In *Proceedings of TREC 2005*.

[5] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC-2005 Enterprise Track. In *Proceedings of TREC 2005*.

[6] N. Craswell, and D. Hawking. Overview of the TREC-2002 Web Track. In *Proceedings of TREC 2002*.

[7] N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins. Panoptic expert: Searching for experts not just for documents. In *Ausweb Poster Proceedings*, 2001.

[8] M. Hertzum and A. M. Pejtersen. The information-seeking practices of engineers: searching for documents as well as for people. *Inf. Process. Manage.*, 36(5):761–778, 2000.

[9] J. Lin. Divergence measures based on the Shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[10] C. Macdonald, B. He, V. Plachouras, and I. Ounis. University of Glasgow at TREC 2005: Experiments in Terabyte and Enterprise tracks with Terrier. In *Proceedings of TREC 2005*.

[11] C. Macdonald and I. Ounis. Voting for candidates: Adapting Data Fusion techniques for an Expert Search task. In *Proceedings of CIKM 2006*.

[12] C. Macdonald and I. Ounis. Using relevance feedback in expert search. In *Proceedings of ECIR 2007*, pages 431–443, LNCS, Springer.

[13] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of OSIR Workshop 2006*, pages 18–25.

[14] D. Petkova and W. B. Croft. Hierarchical language models for expert finding in enterprise corpora. *Proceedings of ICTAI 2006*, pages 599–608.

[15] S. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In *Proceedings of TREC 8*, 2000.

[16] J. J. Rocchio. *Relevance Feedback in Information Retrieval*. Prentice-Hall, 1971.

[17] X. Shen and C. Zhai. Active feedback in ad hoc information retrieval. In *Proceedings of SIGIR 2005*, pages 59–66.

[18] I. Soboroff, A. P. de Vries, and N. Craswell. Overview of the TREC-2006 Enterprise Track. In *Proceedings of TREC 2006*.

[19] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.

[20] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *Proceedings of SIGIR 2005*, pages 512–519.