Expert Search Evaluation by Supporting Documents

Craig Macdonald and Iadh Ounis

Department of Computing Science,
University of Glasgow, Glasgow, G12 8QQ, UK
{craigm,ounis}@dcs.gla.ac.uk

Abstract. An expert search system assists users with their "expertise need" by suggesting people with relevant expertise to their query. Most systems work by ranking documents in response to the query, then ranking the candidates using information from this initial document ranking and known associations between documents and candidates. In this paper, we aim to determine whether we can approximate an evaluation of the expert search system using the underlying document ranking. We evaluate the accuracy of our document ranking evaluation by assessing how closely each measure correlates to the ground truth evaluation of the candidate ranking. Interestingly, we find that improving the underlying ranking of documents does not necessarily result in an improved candidate ranking.

1 Introduction

In large Enterprise settings with vast amounts of digitised information, an *expert search* system aids a user in their "expertise need" by identifying people with relevant expertise to the topic of interest. The retrieval performance of an expert search system is very important. If an expert search system suggests incorrect experts, then this could lead the user to contacting these people inappropriately. Similarly to document IR systems, the accuracy of an expert search system can be measured using the traditional IR evaluation measures such as precision and recall of the suggested candidates. Expert search has been a retrieval task in the Enterprise tracks of the Text REtrieval Conferences (TREC) since 2005 [1], aiming to evaluate expert search approaches.

Most of the existing models for expert search work by examining the set of documents ranked or scored with respect to the query, and then converting this into a ranking of candidates, based on some information about the associations between documents and candidates. However, while various studies have shown that applying known retrieval techniques to improve the quality of the document ranking lead to an improvement in the accuracy of the ranking of candidates [2,3,4], it has not been clear what characteristics in the improved document ranking have caused the increase of retrieval accuracy of the expert search system. This work attempts to approximate an evaluation of the underlying document ranking, to better understand how the document ranking can affect the retrieval accuracy of the expert search system.

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 555–563, 2008.

The objectives of our experiments are two-fold: Firstly, to assess whether the proposed methodology for evaluating the underlying document ranking can produce an accurate estimation of the final accuracy of the expert search system; Secondly, to examine which evaluation measures calculated on the document ranking exhibit the highest correlation with each evaluation measure calculated on the candidate ranking. In doing so, we gain an understanding into how various techniques for expert search behave when the underlying ranking is altered.

The remainder of this paper is as follows: Section 2 briefly reviews several models for expert search, and discusses the evaluation of expert search systems. In Section 3, we show how to approximate an evaluation of the document ranking of an expert search system, and investigate how the document ranking evaluation correlates with the ground truth evaluation of the ranking of candidates. Finally, in Section 4, we provide concluding remarks and points for future work.

2 Models for Expert Search

Given an input list of candidate experts, modern expert search systems work by using documents to form a profile of textual evidence of expertise for each candidate. This associated documentary evidence can take many forms, such as intranet documents, documents or emails authored by the candidates, or web pages visited by the candidate (see [2] for an overview). The candidate profiles can then be used to rank candidates automatically in response to a query.

The most successful models for expert search use an initial ranking or scoring of documents with respect to the query [2,4,5,6]. For instance, in Model 2 of the language models proposed by Balog et al. [5], the probability of a candidate is the sum of the probability of all retrieved documents, multiplied by the degree of association between each document and the candidate. Similarly, in the Voting Model for Expert Search [2], various voting techniques can be applied to aggregate the retrieval scores or ranks of all the retrieved documents associated to each candidate to form the final score for the candidate.

For all these techniques, there are three fundamental parameters that can impact the accuracy of the expert search system: Firstly, the technique(s) applied to generate the underlying ranking of documents impact the final ranking of candidates: various studies have shown that applying techniques (which normally improve a document IR system) improve the 'quality' of the document ranking results in increased accuracy of the candidate ranking [2,3,4]; Secondly, the quality of expertise evidence for each candidate (for instance how documents have been associated to each candidate) has a major impact on the performance of the system [5,7]; Lastly, the manner in which the document evidence is combined for each candidate impacts on how accurate the expert search system is [2].

This work is concerned with the document ranking experimental parameter. While it is possible to evaluate the final ranking of candidates, it has not been possible to determine the properties of a 'high quality' ranking of documents that produces an accurate ranking of candidates, because there has been no direct method of measuring this 'quality'. In the remainder of this section, we review how expert search system evaluation is normally performed, while the next section describes how we can approximate an evaluation of the document ranking.

2.1 Evaluation of Expert Search Systems

The evaluation of expert search systems presents more difficulties than that of a document retrieval system, primarily because a document assessor can read a document, and fairly easily make a judgement as to its relevance. However, an expert search system returns only a list of names, with nothing to allow an assessor to easily determine each person's expertise. To this end, using the TREC paradigm, there are essentially three strategies for expert search system evaluation, to generate relevance assessments for candidates:

Pre-Existing Ground Truth: In this method, queries and relevance assessments are built using a ground truth, which is not explicitly present in the corpus. For example, in the TREC 2005 expert search task, the queries were the names of working groups within the W3C, and participating systems were asked to predict the members of each working group [1]. The problem with this method of evaluation is that it relies on known grouping of candidates, and does not assess the systems for more difficult queries where the vocabulary of the query does not match the name of the working group. Moreover, candidates can have expertise in topics they are not members of working groups on.

Candidate Questionnaires: In this method, each candidate expert in the collection (or a person with suitable knowledge about the candidate experts' expertise areas), is asked if they have expertise in each query topic. While this process can be reduced in size by using pooling of the suggested candidates for each query, the process obviously does not scale to a large collection with hundreds or thousands of candidates. In particular, not all candidates are available to question, or assessors may not have knowledge of every candidates' interests. A derivative of this approach was used to assess the TREC 2007 expert search task in a medium-sized enterprise setting [9].

Supporting Evidence: This last method was proposed for the TREC 2006 expert search task [10]. In this method, each participating system is asked, for each suggested candidate, to provide a selection of ranked documents that supported that candidate's expertise. For evaluation, the top-ranked candidates suggested for each query are pooled, and then for each pooled candidate, the top-ranked supporting documents are pooled. Relevance assessment follows a two-stage process: assessors are asked to read and judge all the pooled supporting documents for a candidate, before making a judgement of his/her relevance to the query. Additionally, the pooled supporting documents which supported their judgement of expertise are marked. Figure 1 shows a section of the TREC 2006 relevance assessments, showing that candidate-0001 has relevant expertise to topic 52. Moreover, a selection of supporting documents are provided, which the relevance assessor used to support that judgement. In the final evaluation, only the

```
52 candidate-0001 2
  52 candidate-0001 lists-015-4893951 2
  52 candidate-0001 lists-015-4908781 2
  ....
52 candidate-0002 0
....
```

Fig. 1. Extract from the relevance assessments of the TREC 2006 expert search task (topic 52). candidate-0001 is judged relevant, with two positive supporting documents shown (lists-015-4893951 etc.). candidate-0002 is not judged relevant.

candidate relevant assessments are used to evaluate the accuracy of the expert search systems.

Once the (candidate) relevance assessments have been generated, using one of the methods described above, it is then simple to evaluate a ranking of candidates using standard retrieval evaluation measures, such as Mean Average Precision (MAP), etc. For clarity, we call these measures Candidate MAP, etc, as they are calculated on the ranking of candidates.

3 Document Ranking Evaluation

As noted above, the current effective models for expert search all take into account the notion of document relevance to the query topic, before ranking the associated candidates. We designate this underlying ranking of documents for the query as $R(Q)$. Because various studies have shown that improving $R(Q)$ has increased the accuracy of the candidate ranking, one could assume that the accuracy of the ranking of candidates is dependant on how well the underlying ranking of documents ranks highly documents related to the relevant candidates.

We aim to approximate an evaluation of the document ranking directly, to aid failure analysis of expert search systems. In doing so, we hope to gain new insights about the desirable characteristics of the document retrieval component of an expert search system, which will help to build more accurate expert search systems. To achieve this approximate evaluation, we use the supporting documents as relevance assessments: a document is assumed to be relevant to a query iff it was judged as a relevant supporting document for a relevant candidate of that query. Then to evaluate the document ranking, we use standard evaluation measures, applied using these supporting document relevance assessments. Mean Average Precision measured on the document ranking is denoted MAP of $R(Q)$.

This work has two central objectives: Firstly, we test if the evaluation using supporting documents of the underlying document ranking can approximate the evaluation of the final candidate ranking; Secondly, to determine which measures calculated on the document ranking best predict various measures calculated on the candidate ranking. For our experiments, we use the set of supporting documents for all relevant candidates from the TREC 2006 expert search task. In particular, 49 queries were assessed, for which there are on average 28.4 candidates with relevant expertise. For each relevant candidate, there is on average

9.8 supporting documents for that judgement, which over all candidates, gives a mean of 134.8 unique supporting documents per query.

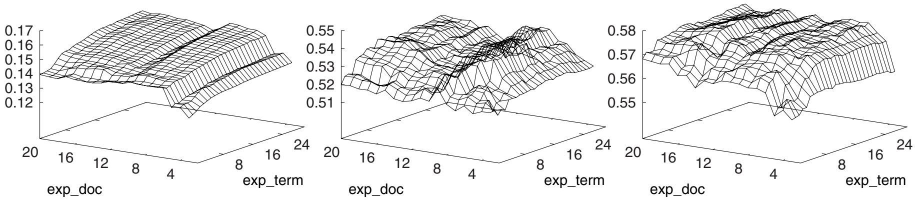
In the following section, we use an expert search system to generate many document rankings and corresponding candidate rankings, and examine how changes in the document rankings are reflected in the candidate rankings. The following section details the experimental setting applied.

3.1 Experimental Setting

The TREC W3C collection is indexed using Terrier [8], removing standard stopwords and applying the first two steps of Porter's stemming algorithm. Documents in the initial ranking $R(Q)$ are ranked using the DLH13 document weighting model [2] from the Divergence from Randomness (DFR) framework. We chose to experiment using DLH13 because it has no term frequency normalisation parameter that requires tuning, and hence, by applying DLH13, we remove the presence of any term frequency normalisation parameter in our experiments. We then create many document rankings by varying the parameters of a document-centric query expansion technique. 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 profile of the candidate. The document candidate associations are not varied, however the applied associations have previously performed robustly on the same task [3].

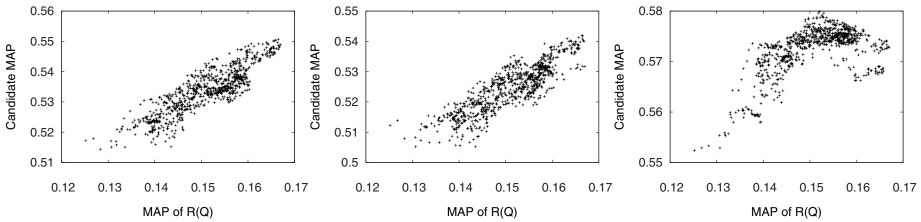
For the combining of document ranking evidence into a ranking of candidates, we use three voting techniques from the Voting Model, namely CombSUM, CombMNZ and expCombMNZ [2], as these provide several distinct methods to transform a document ranking into a candidate ranking. Note that CombSUM is equivalent to the Model 2 approach of Balog et al [5], if a language modelling approach is used to generate $R(Q)$ [3]. For this reason, we do not experiment using the language modelling approach of Balog et al [5].

To assess how the document ranking evaluation correlates with the evaluation of the generated candidate ranking, we need to generate many alternative document rankings for each query, evaluate them, and see how these correlate to the final candidate evaluation measure. To this end, and as mentioned above, we use document-centric query expansion (DocQE) for expert search [3]. In document-centric QE, query expansion is applied on the document ranking, to identify some informative terms from the top-ranked documents (we use the Bo1 DFR term weighting model to measure the informativeness of terms [3]), which are added to the initial query. The expanded query is then re-run to give an enhanced document ranking, which should produce higher retrieval performance when transformed into a ranking of candidates [3]. The number of top retrieved documents to consider (exp_doc) and the number of terms to add to the query (exp_term) are parameters of the query expansion, and by varying these we can generate various initial ranking $R(Q)$ with varying retrieval performances. We vary $1 \leq exp\_term \leq 29$ and $3 \leq exp\_doc \leq 29$, giving 783 different parameter settings.



(a) MAP of R(Q) (b) Candidate MAP (CombSUM) (c) Candidate MAP (expCombMNZ)

Fig. 2. The effect of varying QE parameters (*exp_term* and *exp_doc*) on the various evaluation measures, i.e. MAP on the initial document ranking (denoted MAP of R(Q)), and final candidate MAP calculated on the candidate ranking produced by the CombSUM and expCombMNZ voting techniques. (Note different Z-axis scales).



(a) CombSUM (b) CombMNZ (c) expCombMNZ

Fig. 3. Scatterplots showing the overall correlation between MAP of R(Q) and Candidate MAP, for three voting techniques. (Note different Y-axis scales).

3.2 Document Ranking Correlation

Figure 2(a) shows a surface plot for various settings of the *exp_term* and *exp_doc* QE parameters, evaluated using MAP of $R(Q)$. Secondly, Figures 2(b) & (c) show the retrieval performance achieved when the ranking is aggregated into a ranking of candidates by CombSUM and expCombMNZ respectively[1]. Each point in Figure 2(a), (b) or (c) represents the MAP over the 49 topics in the TREC 2006 expert search task. Comparing these figures we can observe that while the outline of the surfaces between the MAP of R(Q) and candidate MAP plots are similar, the MAP of R(Q) plot is much smoother - this suggests that if the overall correlation trend between MAP of R(Q) and candidate MAP plots is similar, it may be easier for an automated training process (e.g. hill climber or simulated annealing) to train an expert search system on the smoother MAP of R(Q) surface.

Figures 3(a), (b) & (c) show scatterplots of the correlations between MAP of R(Q) vs Candidate MAP for the CombSUM, CombMNZ and expCombMNZ voting techniques respectively. From the figures, it is clear that the accuracy of the voting techniques is dependent on the accuracy of the underlying ranking of documents. In particular, we can quantify this by examining the overall

[1] The plot for CombMNZ is similar to CombSUM, and is hence omitted for brevity.

Table 1. Correlation between various document and candidate ranking evaluation measures, for three voting techniques. The best correlation for each Candidate ranking measure (column) and voting technique are emphasised, while correlations which are statistically different (using a Fisher Z-transform and the two-tailed significance test) from the best correlation ($p < 0.05$) in each column are denoted *.

	Candidate Measures											
R(Q)	CombSUM				CombMNZ				expCombMNZ			
	MAP	MRR	rPrec	P@10	MAP	MRR	rPrec	P@10	MAP	MRR	rPrec	P@10
MAP	**0.8552**	**0.7076**	**0.8192**	0.6461	**0.8561**	**0.6581**	**0.8260**	**0.6661**	**0.4898**	0.0942*	**0.3190**	-0.0397
MRR	0.3503*	0.5008*	0.2889*	0.0492*	0.3031*	0.4151*	0.2868*	0.2274*	0.0570*	-0.2701*	0.2092*	-0.0469
rPrec	0.8256*	0.5737*	0.8086	**0.6959**	0.8519	0.5942*	0.8034	0.6049*	0.4280	**0.2420**	0.1745*	-0.2041*
P@10	0.7225*	0.6008*	0.6955*	0.5300*	0.7340*	0.5378*	0.7206*	0.5929*	0.4235	0.0665*	0.0361*	**0.0361**

correlation between the ranking of settings by MAP of R(Q) and the Candidate MAP, using Spearman's ρ. In these cases, $\rho = 0.8552$ and $\rho = 0.8561$ over the 783 points each, for CombSUM and CombMNZ respectively. For expCombMNZ, which performs better overall, the correlation is lower ($\rho = 0.4898$), and interestingly a 'tail-off' in Candidate MAP can be observed for MAP of R(Q) > 0.15. Indeed, this technique exhibits a rather unexpected trait in the sense that improving the document ranking does not always result in an improved candidate ranking accuracy. We suspect that this is an example of a form of over-fitting of the QE technique to the document ranking evaluation. In general, we conclude that to improve the accuracy of an expert search system, we can apply techniques that are known to improve the accuracy of a standard document retrieval system, however, some techniques (e.g. expCombMNZ) can suffer when the document ranking is over-fitted to the R(Q) evaluation, and thus require further investigation to fully understand this phenomenon.

Next, we investigate which measures calculated on the initial document ranking predict best various evaluation measures for the candidate ranking. In doing so, we aim to understand what characteristics in the document ranking affect the generated candidate ranking. Table 1 presents the Spearman's ρ correlation between various evaluation measures on the document ranking (R(Q)), and the final ranking of candidates, for the CombSUM, CombMNZ and expCombMNZ voting techniques. The evaluation measures applied are MAP, precision at R documents (rPrec), reciprocal rank of first relevant document (MRR) and precision @10 (P@10).

From the results, we can draw the following conclusions: MAP and rPrec on the document ranking are good predictors for both the candidate MAP and rPrec measures. This is not surprising, given that rPrec is often the most correlated measure to MAP [11]. In general, for CombSUM and CombMNZ, MAP of R(Q) is the best predictor for any candidate ranking measure (an exception is CombSUM, where rPrec is a slightly better predictor for P@10). This is intuitive, as the voting techniques investigated here are recall orientated - i.e. they examine all the retrieved document associated with each candidate, so it makes sense that even small changes lower down the document ranking improve the overall effectiveness of the voting technique. In contrast, despite the higher retrieval performance of expCombMNZ technique, lower correlations are observed. In particular, MAP and rPrec on R(Q) are the best predictors for candidate MAP. P@10 is also a good predictor, due to the natural focus of expCombMNZ on the top of the document ranking. However, it appears to be impossible to predict the candidate P@10 measure for expCombMNZ, which is unexpected, because MAP and P@10 are normally strongly correlated [11].

Overall, while in general, we conclude that in order to improve an expert search system, it appears to be most effective to apply retrieval techniques that improve MAP, regardless of the evaluation measure that it is desired to improve.

4 Conclusions

The current effective expert search models all take into account, in some way, the relevance score of the documents with respect to the query, which are then converted into a ranking of candidates. Moreover, previous works on expert search show that somehow improving the quality of the underlying ranking of documents ($R(Q)$) results in a more accuracy expert search system.

In this work, we have proposed an approximate evaluation of $R(Q)$ using the supporting documents as relevance assessments. In our experiments, we examined how closely the R(Q) evaluation correlates to the final candidate ranking, using various evaluation measures, across various input document rankings of varying quality. Our experiments found that the document ranking could be evaluated using the proposed methodology. Furthermore, while various measures can be used to measure the quality of R(Q), for the voting techniques applied, MAP appears to be the most effective predictor of the candidate evaluation measures.

The initial step taken in this work towards the evaluation of expert search systems using the document ranking is important as the current evaluation is awkward due to its second-order nature. By showing that the accuracy of the ranking of candidates generated by an expert search system is indeed linked to the quality of the underlying document ranking, failure analysis becomes easier. Moreover, we are able to gain more insights into the characteristics of the document ranking which influence the generated candidate ranking.

In this paper, we did not evaluate the document ranking with real document relevance assessments, instead approximating these using the supporting document as relevance assessments. The newly available TREC 2007 Expert Search test collection [9] is the natural next step for this work, as it contains relevance assessments for candidates and documents on the same query topics. Additionally, using a more diverse source of document rankings than varying query expansion parameters would allow a fuller understanding of the evaluation methodology.

References

1. Craswell, N., de Vries, A.P., Soboroff, I.: Overview of the TREC 2005 Enterprise Track. In: Proceedings of TREC 2005, Gaithersburg, MD (2006)
2. Macdonald, C., Ounis, I.: Voting for candidates: Adapting Data Fusion techniques for an Expert Search task. In: Proceedings of ACM CIKM 2006, Arlington, VA (2006)

3. Macdonald, C., Ounis, I.: Using Relevance Feedback in Expert Search. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 431–443. Springer, Heidelberg (2007)
4. Petkova, D., Croft, W.B.: Hierarchical language models for expert finding in enterprise corpora. In: Proceedings of ICTAI 2006, pp. 599–608 (2006)
5. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. In: Proceedings of ACM SIGIR 2006, Seattle, WA, pp. 43–50 (2006)
6. Cao, Y., Li, H., Liu, J., Bao, S.: Research on Expert Search at Enterprise Track of TREC 2005. In: Proceedings of TREC 2005, Gaithersburg, MD (2006)
7. Macdonald, C., Ounis, I.: High Quality Expertise Evidence for Expert Search. In: Macdonald, C., et al. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 283–295. Springer, Heidelberg (2008)
8. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: Proceedings of OSIR Workshop 2006, Seattle, WA (2006)
9. Bailey, P., Craswell, N., de Vries, A.P., Soboroff, I.: Overview of the TREC-2007 Enterprise Track. In: Proceedings of TREC-2007, Gaithersburg, MD (2008)
10. Soboroff, I., de Vries, A.P., Craswell, N.: Overview of the TREC-2006 Enterprise Track. In: Proceedings of TREC 2006, Gaithersburg, MD (2007)
11. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proceedings of ACM SIGIR 2004, Sheffield, UK, pp. 25–32 (2004)