The Influence of the Document Ranking in Expert Search

Craig Macdonald, Iadh Ounis
Department of Computing Science
University of Glasgow, Scotland, UK
{craigm,ounis}@dcs.gla.ac.uk

ABSTRACT

The retrieval effectiveness of the underlying document search component of an expert search engine can have an important impact on the effectiveness of the generated expert search results. In this large-scale study, we perform novel experiments in the context of the document search and expert search tasks of the TREC Enterprise track, to measure the influence that the performance of the document ranking has on the ranking of candidate experts. In particular, we show, using real and simulated document rankings, that while the expert search system performance is related to the relevance of the retrieved documents, surprisingly, it is not always the case that increasing document search effectiveness causes an increase in expert search performance.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Performance, Experimentation

Keywords: Expert Search, Document Search

1. INTRODUCTION

Many models for expert search are based on the premise that the more a document is related to the topic of the query, the more likely that candidates associated to that document will have relevant expertise to the query [2, 3]. However, the manner in which the strength of topicality is weighted – how much each document is related to the query – has seen less analytical research. Typically, experiments have shown that by applying a known information retrieval (IR) technique which usually improves the retrieval performance of a document search engine, performance is also improved for the expert search engine (for example [2, 3]).

What remains unclear from these analyses is which documents in a document ranking are actually useful for producing an accurate ranking of experts. Should the document search component be trained to give as many relevant documents as possible, or only to highly rank a few key pages for the topic (i.e. focusing on recall or precision)? The aim of this paper is to revisit the document search component, by analysing various aspects of the quality of the document ranking to ascertain how these affect the retrieval performance of the expert search system.

CIKM'09, November 2–6, 2009, Hong Kong, China.

The central contribution of this work is a novel large-scale study – using distinct sources of document rankings – of which factors of the quality of the document ranking affect the retrieval performance of a range of expert search approaches. For instance, our methodology allows us to determine if a particular expert search approach "prefers" a document ranking of high precision more so than another approach. In our study, we use two diverse sources of document rankings, namely the document rankings submitted by participants to the TREC 2007 document search task, and fictitious rankings with various simulated retrieval performances. Surprisingly, we find that it is not always the case that increasing document search effectiveness causes an increase in expert search performance.

2. EXPERT SEARCH APPROACHES

Various approaches have been proposed for expert search which use documentary evidence of expertise for each candidate (called candidate profiles) to rank candidates in response to a query. In general, the most effective approaches are based on mapping a ranking of documents into a ranking of candidates. This is the approach taken by the Voting Model [3], which sees the expert search task as a voting process. In the Voting Model, the ranking of documents (denoted $R(Q)$) defines votes for candidates to be retrieved: each time a document associated with a candidate is ranked in $R(Q)$, then this is an implicit vote for that candidate to have relevant expertise to the query. The so-called *Model 2* approach works in a similar manner [2], but, unlike the Voting Model, is limited to use in language modelling settings.

The Voting Model defines many voting techniques, each corresponding to a different way of aggregating the votes from a ranking of documents into a ranking of candidates. Using the Voting Model, we have the advantage of experimenting with various voting techniques, each of which encapsulates different intuitions about how evidence from the document ranking is used to rank experts. In this work, we study six voting techniques (summarised in Table 1), each of which uses either the score of a document with respect to the query, or the rank of a document in the underlying ranking. $profile(C)$ defines the set of documents associated to each candidate as evidence of their expertise.

3. IMPROVING EXPERT SEARCH PERFORMANCE

It seems intuitive that a more refined, *higher quality* document search component will allow an expert search system to attain improved retrieval performance. For example, training the document search component [2, Ch. 4][3, Ch. 5], or applying field-based weighting models or query term

Name	Relevance score of candidate is:
ApprovalVotes	$\|\mathrm{D(C,Q)}\|$
RecipRank	sum of inverse of ranks of docs in $D(C,Q)$
BordaFuse	sum of ($\|R(Q)\|$ - ranks of docs in $D(C,Q)$)
CombMAX	maximum of scores of docs in $D(C,Q)$
expCombSUM	sum of exp of scores of docs in $D(C,Q)$
expCombMNZ	$\|D(C,Q)\| \times$ expCombSUM

Table 1: Summary of voting techniques used in this paper. $D(C,Q)$ is the set of documents $R(Q) \cap profile(C)$. $\|\cdot\|$ is the size of the described set.

proximity [3, Ch. 6] have been shown to improve an expert search engine. Different formulations of query expansion on the document ranking have both been shown to help or hinder expert search performance [3, Ch. 7]. Moreover, the application of Web IR features on the document ranking (e.g number of inlinks, URL length) was found not to be as useful as other expert search-specific evidence, such as the proximity of query terms to occurrences of candidates' names [3, Ch. 7].

In these previous works, the application of different techniques has improved the document ranking in some way that has often resulted in an improved candidate ranking. However, the aspects of the document ranking which had an impact on the expert search performance are unknown. Moreover, there were no relevance assessments with which to directly evaluate the document ranking in context. To tackle this, in [4], we studied approximating a document ranking evaluation. In contrast, in this study, we use many document rankings with known retrieval performances as input to various voting techniques, and compare and contrast the document ranking performance with the corresponding expert search performance. In other words, we are testing whether topically relevant documents are necessary and sufficient expertise evidence.

4. COMPARING DOCUMENT SEARCH & EXPERT SEARCH PERFORMANCE

In this work, we aim to address the following question: when used as input to an expert search approach, which aspects of a document ranking have an impact on the retrieval performance of the generated candidate ranking? The particular document ranking aspects which produce accurate candidate retrieval performance may depend on the particular voting technique applied. For example, for effective performance, ApprovalVotes requires many documents that are related to the topic and associated to relevant candidates to be retrieved, while minimising the number of retrieved documents associated to irrelevant candidates. In contrast, for other voting techniques such as RecipRank or BordaFuse, the document ranking should highly rank documents that are related to the topic and associated to relevant candidates. Documents that are not about the topic or associated to irrelevant candidates should not be retrieved, or should be ranked as lowly as possible; RecipRank focuses more on the top of the document ranking than BordaFuse.

The difficulty in measuring the quality of the document ranking is that there are no measures which easily encapsulate these preferences of the various voting techniques on the document ranking. Instead, we examine both the effectiveness of the document ranking when used for a document retrieval task, and the retrieval performance of the ranking of candidates generated by use of a voting technique on such a document ranking. By comparing the performance of the document ranking to the accuracy of the generated ranking of candidates, we aim to draw conclusions about the features of the document ranking which matter most for a given voting technique.

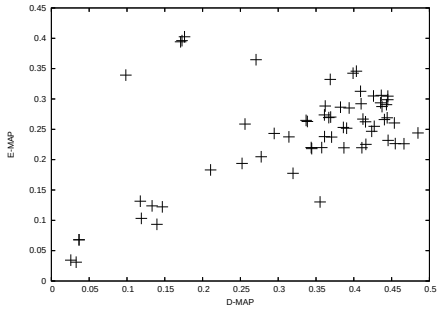

Figure 1: Scatter plot showing the correlation between D-MAP & E-MAP for the BordaFuse voting technique.

In particular, we use the document search task and the expert search task of the TREC 2007 Enterprise track. Importantly, for both tasks, participants used a common set of queries, and a common document collection called CERC. In the document search task, systems should identify relevant documents for each query, while for the expert search tasks, relevant experts should be suggested.

In the following sections, we aim to determine how the retrieval performance of an IR system on the document search task has an impact on the accuracy of the generated ranking of candidates, when that IR system is used as input to a given voting technique. We perform this experiment using two sources of rankings, namely the TREC 2007 submitted runs (Section 4.1) and simulated IR systems (Section 4.2).

4.1 Real Document Rankings

We are interested in determining how document rankings, of various but known retrieval performances, affect the performance of various voting techniques. To achieve this, we measure the performance of 63 real document rankings which were actual submitted runs to the TREC 2007 document search task, and then compare with the performance of each when used as the input to a voting technique. The relevance assessments of the TREC 2007 document search task (DS07) are used to assess the quality of the document rankings, while the relevance assessments of the TREC 2007 expert search task (EX07) are used to measure the accuracy of the generated candidate rankings. For clarity, the evaluation of a document ranking using MAP with DS07 judgements is denoted D-MAP, while evaluation of a candidate ranking is denoted E-MAP.

The associations between candidates and documents (the candidate profiles) form the most important experimental parameter. To identify possible candidates in the collection, we search the documents for email addresses of the form firstname.lastname@csiro.au. To generate the document-candidate associations for each candidate, documents are identified by the presence of the candidate's exact name or email address [2, Ch. 7][3, Ch. 6], which has been shown to be effective for this task in the past.

Figure 1 illustrates D-MAP vs. E-MAP over all submitted document search runs, when applied using the BordaFuse voting technique. We observe that while there are some outliers, we can see that there is a rough correlation between D-MAP and E-MAP. A higher D-MAP makes BordaFuse more

likely to have a higher E-MAP. However, around the range of D-MAP 0.28–0.45, there is less correlation, and we have a less clear picture. We note that of the runs with D-MAP in this range, when applied to BordaFuse, some perform stronger than others. This means that the exact characteristics of the document ranking desired by BordaFuse are not being well measured by D-MAP – of the outliers, there are some runs with low D-MAP but with strong E-MAP. On further inspection, we found that these runs have returned far less documents than the other runs. This degrades their D-MAP performance, however (E-)MAP on the EX07 task is improved by considering less documents in the document ranking [3, Ch. 6].

We can quantify the extent to which the system rankings by D-MAP and E-MAP in Figure 1 are correlated, using the Spearman's rank correlation co-efficient ρ. Moreover, because it has been previously noted that the voting techniques performed best for the EX07 task using only the top 50-ranked documents, we perform our correlation experiments where the $R(Q)$ for every query has been cutoff after 50 retrieved documents.

The top part of Table 2 presents the correlations between various document search task measures and the accuracy of various voting techniques. We assess the D-MAP, D-MRR, D-nDCG[1], D-P@10 and D-Recall measures, to determine which are correlated with the official measures of the expert search task, namely E-MAP and E-MRR. The best correlations for each candidate ranking measure and voting technique (row) are emphasised, while correlations which are statistically different (using a Fisher Z-transform and the two-tailed significance test) from the best correlation in each row are denoted * ($p < 0.05$) and ** ($p < 0.01$). Finally, the best E-MAP and E-MRR performances for each voting technique for any input document ranking are also reported.

From the results in Table 2, we observe overall strong positive correlations, suggesting that the performance of various voting techniques can be predicted by various measures calculated on the document ranking. However, from the overall trends it is not the case that for each E-measure, the corresponding D-measure is the most correlated. Instead, various voting techniques focus on different parts of the document ranking in different ways, and the document ranking quality affects their overall accuracy in different ways. In the following, we detail how document ranking quality affects each voting technique in turn.

ApprovalVotes: Highest correlations are with D-Recall. This is expected, as this technique only considers the number of votes, which we postulate will be highly correlated with D-Recall. Other measures which examine the entire ranking, e.g. D-MAP, D-nDCG, and D-P@50 are also strongly correlated with E-MRR and, in particular, E-MAP. Conversely, less strong correlations are observed with measures that examine only the higher ranked documents (e.g. D-MRR or D-P@10), which is expected, as ApprovalVotes treats all retrieved documents equally, regardless of rank.

BordaFuse: High correlations with D-MAP, D-nDCG & D-Recall, showing that while BordaFuse uses all the retrieved documents, it focuses on the more highly ranked ones. The higher correlation for D-nDCG than D-MAP indicates candidate ranking performance is enhanced by a document ranking which ranks highly relevant documents before relevant ones.

RecipRank: The trends exhibited by RecipRank are similar to BordaFuse, however with slightly less high correlations overall. Surprisingly, there is no bias toward top-heavy D-measures such as P@10.

CombMAX: Intuitively, CombMAX is most influenced by the top of the document ranking, hence it is expected that a retrieval system which has good success at early ranks will likely enable CombMAX to perform well, explaining why CombMAX only shows high correlations with D-MRR.

expCombSUM: Similarly to BordaFuse, we find that expCombSUM has a high correlation with D-MAP and D-nDCG, showing a focus towards the top of the document ranking (particularly highly relevant documents). The correlations for D-Recall is only slightly higher than D-nDCG, and not significantly so.

expCombMNZ: expCombMNZ also exhibits high correlations with D-MAP, D-nDCG, & D-Recall. Compared with expCombSUM, D-Recall is relatively more important than D-MAP, which is explained by the number of votes component in expCombMNZ.

Overall, the strength of the correlations exhibited are promising, indicating that there is a strong likelihood of a relationship between the topical retrieval performance of $R(Q)$, and the performance of a voting technique. In particular, our intuitions about the "preferences" of the voting techniques are confirmed - e.g. CombMAX prefers a high precision ranking. When choosing a voting technique, a system designer should choose one which has a high correlation to a document ranking measure on which the existing document IR system is particularly effective. For example, a document IR system which has good MRR should use CombMAX, while another with high Recall/MAP may use expCombSUM or expCombMNZ.

However, we do not find any 100% correlations, showing that not every improvement in document search effectiveness can have a positive impact on an expert search engine. The correlations found here do not show that topic relevance document retrieval performance is perfectly related to candidate retrieval performance. This infers that there are some characteristics of the document ranking which are important to the voting techniques that are not being captured by the topical relevance document evaluation measures.

Recall that the majority of the real document rankings had a D-MAP between 0.28 and 0.45. Given these correlations, another natural question that arises is whether the observed correlations hold for a larger range of possible D-MAP values. In the next section, we use simulation to generate document rankings of various document retrieval performances, and determine how effective these are for expert search using the considered voting techniques.

4.2 Simulated Document Rankings

So far, we have been investigating how real document rankings of various retrieval effectiveness affect the expertise retrieval performance when applied to various voting techniques. We now extend our experiments to use simulated document rankings, which cover an extended range of possible D-MAP values. We use the AP simulation algorithm proposed by Turpin & Scholer in [6], which makes improving or degrading random swaps of relevant and irrelevant documents until the target AP performance is achieved (or no more swaps are possible).

Firstly, for each query, the D-MAP range is split into 20 equal-sized bins (size 0.05). Then, we generate 20 rankings in each bin, using a random target D-MAP value within the range of the bin, to give a total of 400 simulated "runs". Each run, which has 50 queries of very similar effectiveness,

[1]DS07 task has ternary judgments [1].

Voting Technique	Expert Measure		ρ Correlation Values (by Document Search Measure)						
	Name	Max	D-MAP	D-nDCG	D-MRR	D-P@10	D-P@30	D-P@50	D-Recall
Real Rankings									
ApprovalVotes	E-MAP	0.4773	0.7318	0.7633	0.3848**	0.6570	0.7497	0.7598	**0.7915**
	E-MRR	0.6174	0.6497	0.6749	0.3439**	0.5732	0.6468	0.6751	**0.7023**
BordaFuse	E-MAP	0.4860	0.8292	0.8584	0.4808**	0.7760	0.8341	0.8252	**0.8650**
	E-MRR	0.6243	0.8216	0.8392	0.4439**	0.7517	0.8015	0.7882	**0.8425**
RecipRank	E-MAP	0.4893	0.7728	0.8091	0.4376**	0.7140	0.7762	0.7759	**0.8128**
	E-MRR	0.6262	0.7254	0.7533	0.3986**	0.6523	0.7166	0.7277	**0.7594**
CombMAX	E-MAP	0.4829	0.1390**	0.1988**	**0.5878**	0.3113	0.1884**	0.1374**	0.1602**
	E-MRR	0.6187	0.0601**	0.1261**	**0.5806**	0.2436*	0.1172**	0.0685**	0.0893**
expCombSUM	E-MAP	0.4961	0.6914	0.6917	0.2245**	0.6232	0.6722	0.6482	**0.7196**
	E-MRR	0.6647	0.6639	0.6719	0.2565**	0.6129	0.6477	0.6223	**0.7012**
expCombMNZ	E-MAP	0.4956	0.6714	0.6750	0.2197**	0.5996	0.6406	0.6119	**0.7008**
	E-MRR	0.6539	0.6749	0.6896	0.3072**	0.6382	0.6522	0.6201	**0.7064**
Simulated Rankings									
ApprovalVotes	E-MAP	0.3051	0.9161	0.9174	**0.9234**	0.9194	0.9165	0.9160	0.9162
	E-MRR	0.4123	0.8726	0.8739	**0.8826**	0.8769	0.8728	0.8724	0.8729
BordaFuse	E-MAP	0.3147	0.8990	0.9004	**0.9089**	0.9007	0.8998	0.8989	0.8987
	E-MRR	0.4295	0.8569	0.8582	**0.8667**	0.8582	0.8575	0.8568	0.8564
RecipRank	E-MAP	0.3083	0.9125	0.9144	**0.9212**	0.9155	0.9132	0.9123	0.9127
	E-MRR	0.4178	0.8583	0.8603	**0.8690**	0.8625	0.8587	0.8580	0.8585

Table 2: Correlations (Spearmans's ρ) between the expert search performance of various voting techniques, compared to the retrieval performance of rankings from the 63 real document search task runs, and the 400 simulated rankings. The best achieved value for each expert search evaluation measure is also shown.

is then used as input to a voting technique. As the simulation does not produce document relevance scores, we focus only on rank-based voting techniques in these experiments. Moreover, each document ranking is unique, using a different ordering of the relevant and irrelevant documents, which may have an impact on the effectiveness of the used voting techniques that consider the ordering of documents.

The second part of Table 2 presents the correlations between various document search task measures for the simulated retrieval systems and the accuracy of three voting techniques using them. On comparing these results with those from the top part of Table 2, we note considerably stronger correlations. This reinforces, that across a full range of possible document search performance, there appears to be a link between the overall topic relevance quality of a document IR system, and its likelihood to be useful as a component of an expert search engine.

However, in contrast to our earlier correlation results, we note that all voting techniques are mostly correlated with D-MRR, and that there are no significant differences between the correlation measures. On further inspection of the simulated document rankings, we found that all D-measures were very similar (e.g. D-MAP vs. D-MRR has $\rho = 0.9042$ for the 400 simulated document rankings, compared to $\rho = 0.4134$ for the 63 real document rankings). Future work will examine how to create more realistic simulated rankings which have different performances on different queries, perhaps starting from the real document rankings.

Table 2 also presents the maximum E-MAP and E-MRR values achieved for each voting technique by any document ranking. From these, we note that the maximum achieved E-MAP and E-MRR values using the simulated rankings are not as high as those from the real TREC runs, even though the simulation experiments contain systems with almost perfect D-MAP document rankings. For instance, the highest E-MAP (0.3147, BordaFuse) was produced by a document ranking with a D-MAP of only 0.6590. These results strengthen those reported in [5], which postulates that not all relevant on-topic documents may be good indicators of expertise evidence, and their exact ordering has an impact on the retrieval performance achievable by a voting technique. It is also possible that there exist some irrelevant documents which can be of benefit to an expert search voting technique [5], and for a ranking with very good D-MAP, these documents have been suppressed, to the detriment of expert search effectiveness.

5. CONCLUSIONS

This work is the first large-scale empirical study into the influence of the document ranking in an expert search system. In particular, we studied this influence on several voting techniques from the Voting Model. However, the results here should generalise to other expert search approaches such as [2]. We experimented with both real and simulated document rankings, and showed that there is a correlation between the ability of the document ranking system to retrieve topically relevant documents with the ability of voting techniques to retrieve an accurate ranking of candidate experts. In particular, using the real document rankings, D-MAP, D-nDCG and D-Recall were all shown to be important predictors of expert search performance. However, from the low maximal performances using the simulated rankings, it is clear that increasing the quality of the input document ranking does not always result in an increase in the retrieval performance of the resulting ranking of candidates. Future work will investigate more advanced simulations, possibly using real document rankings.

6. REFERENCES

[1] P. Bailey, N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC-2007 enterprise track. In *Proceedings of TREC-2007*.

[2] K. Balog. *People search in the enterprise*. PhD thesis, Univ. of Amsterdam, 2008.

[3] C. Macdonald. *The voting model for people search*. PhD thesis, Univ. of Glasgow, 2009.

[4] C. Macdonald and I. Ounis. Expert search evaluation by supporting documents. In *Proceedings of ECIR-2008*, pp 555–563.

[5] C. Macdonald and I. Ounis. On perfect document rankings for expert search. In *Proceedings of SIGIR-2009*, pp 740–741.

[6] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of SIGIR-2006*, pp 11–18.