Ranking expansion terms with partial and ostensive evidence

Ian Ruthven	Mounia Lalmas	Keith van Rijsbergen
Department of Computer and Information Sciences	Department of Computer Science	Department of Computing Science
University of Strathclyde	Queen Mary University of London	University of Glasgow
Glasgow G1 1XH	London, E1 4NS	Glasgow G12 8QQ
Ian.Ruthven@cis.strath.ac.uk	mounia@dcs.qmul.ac.uk	keith@dcs.gla.ac.uk

Abstract

In this paper we examine the problem of ranking candidate expansion terms for query expansion. We show, by an extension to the traditional F_4 scheme, how partial relevance assessments (*how* relevant a document is) and ostensive evidence (*when* a document was assessed relevant*)* can be incorporated into a term ranking function. We then investigate this new term ranking function in three user experiments, examining the performance of our function for automatic and interactive query expansion. We show that the new function not only suggests terms that are preferred by searchers but suggests terms that can lead to more use of expansion terms.

1. INTRODUCTION

Information retrieval (IR) systems are intended to retrieve documents that are relevant to a searcher's information need, usually expressed as a query. However selecting good words (or *terms*) to use as a query can be difficult. If a searcher has found some relevant material she can avoid generating more query terms by asking the system to suggest or add terms to her query. This process is generally known as Relevance Feedback (**RF**) (Harman, 1992): the system exploits those documents the searcher considered relevant to create a better representation of the searcher's information need.

RF is generally composed of three stages; the system first selects possible candidate terms[1] to add to the query and ranks these terms according to some measure of how useful the terms might be in a new query (*term ranking*), the system then selects a number of these terms to add to the query (*query expansion*), and finally the system weights the terms before carrying out a new retrieval (*term weighting*).

In this paper we concentrate on the first stage – term ranking – deciding which terms are most likely to be useful in a new query. The reason that this stage is important is that most RF applications will only choose a small proportion of the candidate expansion term to add to the query. This is not only more computationally efficient than adding all candidate expansion terms (Salton &

Buckley, 1990), but the retrieval effectiveness of a small set of good terms is usually as good as, (Salton & Buckley, 1990), or better than, (Harman, 1992), adding all candidate expansion terms. In addition, adding relatively few expansion terms means that the searcher can easily edit the reformulated query manually.

Standard methods of ranking terms, e.g. F_4, Porter's scheme, or *wpq* (surveyed in Efthimiadis, 1995), treat all relevant documents as a uniform set; all documents are treated as being of equal relevance and no attention is paid to *when* in the search the documents were assessed relevant. These techniques, then, do not incorporate important aspects of searching such as the degree to which a document is relevant to a searcher, (Spink, Greisdorf and Bateman, 1998), or the temporal nature of relevance, (Vakkari, 2000a).

The term ranking function we present in this paper incorporates the non-binary nature of relevance (through the use of partial relevance assessments) and the temporal nature of information seeking (through the use of ostensive evidence). These are to be discussed, along with the motivation and methodology for incorporating these aspects, in section 2. This is followed by a description of a user study on the term ranking function. In section 3, we give a brief introduction to the overall experimental system used in our experiments, and, in section 4, we present the experimental methodology used. In sections 5 –7 we present the experiments we performed. We conclude with a discussion in section 8.

2. TERM RANKING FUNCTION

Our intention is to show that traditional RF techniques can be extended to incorporate more realistic assumptions about searching. Most RF algorithms perform statistical analyses of what searchers assess as relevant: the content of the documents marked relevant by searchers. However, these RF algorithms typically do not consider the complexity behind the process of making relevance assessments. As noted above RF algorithms usually assume binary relevance and assume that a searcher's definition of relevance does not change over the course of a search. However many studies of how searchers assess documents show that relevance assessments can be relative to each other, e.g. (Florance and Marchionini, 1995, Tiamiyu and Ajiferuke, 1988), dynamic, e.g. (Vakkari 2000a, 2000b) and dependent on individual features such as task, and domain knowledge, e.g. (Heuer, 1999). The process of assessing relevance is, therefore, a complex process. However the output of this complex process – the relevant assessments – are compressed into a simple representation for use by RF algorithms – a set of relevant documents.

What motivated us in this study is the belief that advanced search engines require to take more notice and use of evidence from how searchers are interacting with the system. In particular the evidence provided by searchers whilst interacting should be combined with the retrieval algorithms themselves to provide integrated search systems.

Our term ranking function is one example of this, and is based on the standard F_4 term weighting function (Robertson and Sparck Jones, 1976). Although this function was specifically designed to weight query terms based on relevance information, it has been heavily investigated as a means of ranking terms for query expansion, (Efthimiadis, 1995).

The F_4 function, Equation 1, is based on the odds of how likely term *t* is to appear in a relevant document to how likely term *t* is to appear in a non-relevant document. The higher the F_4 weight for term *t* the more likely that *t* appears in one of the relevant documents used to calculate the F_4 weight. To order terms for query expansion, all candidate expansion terms are assigned an F_4 weight and ranked in decreasing order of their F_4 weight.

$$F_4(t) = \ln\left(\frac{r_t(N - n_t - R + r_t)}{(n_t - r_t)(R - r_t)}\right)$$

Equation 1: F_4 term weighting function

where r_t is the number of relevant documents containing term *t*, *R* is the number of relevant documents found so far, n_t is the number of documents containing term *t* and *N* is the number of documents in the collection.

We propose a new term ranking algorithm, the F_4_*po* algorithm[2], Equation 2, that is composed of two components; one component that measures the information coming from partial relevance assessments, section 2.1, and one component that calculates the ostensive evidence for a term, section 2.2. Although there are other possible methods for combining these two components we have made these two components separate, to allow for a future separate study of the effect of the two components.

$$F_4\_po_t = partial_t * ostensive_t$$

Equation 2: F_4_*po* term ranking scheme

2.1 Incorporating partial relevance assessments - $partial_t$ component

The use of partial relevance assessments – allowing searchers to make non-binary assessments on the relevance of retrieved documents has long been

seen as important in obtaining more accurate and realistic assessments of a document's relevance to a searcher, e.g. (Borlund, 2000, Spink et al., 1998).

In the majority of RF interfaces, searchers are asked to make assessments on entire documents as individual entities. This infers that searchers make assessments on the complete document and can identify relevant material. However, often only part of a document may be relevant and the criteria for relevance themselves may be vague, i.e. the searcher may not yet have a well-defined idea of what information is actually required. These two issues point to the partiality of relevance. The latter definition of partial relevance – the vagueness of the searcher's criteria for relevance – has been explored by Spink et al. (Spink et al, 1998) who show a correlation between the number of partial relevance assessments and how well-defined was a searcher's information need. Vakkari (Vakkari 2000b) also showed that a searcher's lack of understanding of their search task correlated with a high number of partial relevance assessments. Incorporating some measure of the *degree* of relevance into the RF process is therefore important in modelling what may be of interest to a searcher.

Our interface (section 3, Figure 2) asks searchers to indicate, using a scroll bar, how relevant is an individual document to their search. Internal to the system this is mapped to a number between 1 and 10, with 0 indicating non-relevant. In our term ranking function we treat these partial relevance assessments as part of a complete relevance assessment, e.g. a document that is assigned a relevance score of 10 by the searcher is treated as a complete relevant document, whereas a document that received a relevance score of 5 is treated as half a relevant document, and so on.

This is integrated into the $partial_t$ component of our algorithm by replacement of the variables r_t, R, n_t and N in the original F_4 weight in the following manner: r_t is the sum of all relevance scores for relevant documents containing term t, R is the sum of all relevance scores for all relevant documents, n_t and N are replaced by n_t*10, N*10, respectively[3]. This means that the higher the relevance scores for documents containing term t the higher the $partial_t$ score for term t.

2.2 Incorporating ostensive evidence - $ostensive_t$ component

The previous section was motivated by the argument that partial relevance assessments are necessary to capture degree of relevance of a document to a searcher. However, the relevance of a document is subject to change throughout the course of a search. This may happen as the result of searchers changing their criteria for relevance or, as indicated by Vakkari (Vakkari,

2000b), the searcher either developing more knowledge about the task or requiring different types of information at different stages in a search.

Although it is difficult to establish *why* a searcher may have made a particular relevance decision, we can allow for the possible change in relevance criteria by the use of *ostensive* evidence. (Campbell and Van Rijsbergen, 1996) argued that *when* in a search a document was marked relevant should be treated as important. Therefore the documents most recently marked relevant are more indicative of what the searcher currently finds relevant – provide more ostensive evidence as to relevance. The incorporation of ostensive evidence, then, does not target reasons for relevance but asserts that newly assessed documents are more likely to demonstrate the searcher's current criteria for relevance.

The ostensive evidence for a term is given by Equation 3.

$$ostensive_t = \left(\sum_{j=1}^{s} j * r_{jt} \right) \Big/ \max_{ostensive}$$

Equation 3: Calculation of $ostensive_t$ component

where s = total number of feedback iterations, r_{jt} = number of relevant documents containing term t in iteration j , $\max_{ostensive}$ = maximum possible ostensive evidence

In Equation 3 the ostensive weight of term *t*, is based on a proportion of the ostensive evidence for *t* relative to the maximum ostensive weight that could be assigned to a term, $\max_{ostensive}$. This maximum ostensive weight will be equal to 1, if all relevant documents, at every iteration of feedback, contained the term *t*. The ostensive evidence for term *t* is the sum of the relevant documents containing *t* multiplied by the iteration in which the documents were marked relevant. Therefore the more relevant documents term *t* appears in, the higher weight it receives and the more recently-viewed relevant documents *t* appears in the higher weight it receives. What the $ostensive_t$ component measures then is how indicative of relevance term *t* is at the current search stage.

An example of this is shown in Figure 1, for two terms – term *t* and term *s*, based on the data given in Table 1. In Table 1, we have 5 iterations of feedback. At each iteration a number of documents are marked relevant (**R** row 5), some of which contain term *t*, ($\mathbf{r}_t$ row 3), and some of which contain term *s* ($\mathbf{r}_s$ row 4). The maximum ostensive weight for both terms is identical: both terms could have appeared in all relevant documents at each iteration of feedback. What differs between the two terms is when the documents containing the terms were marked relevant: the relevant documents containing

term *t* were assessed as relevant later in the search than the relevant documents containing term *s*. Hence term *t* receives a higher ostensive weight than term *s*.

Iterations of feedback						
	1	**2**	**3**	**4**	**5**	**Total**
$\mathbf{r}_t$	1	0	0	1	5	7
$\mathbf{r}_s$	5	1	0	0	1	7
R	5	2	3	1	10	21

Table 1: Example ostensive data

$$\max_{ostensive} = (5*1) + (2*2) + (3*3) + (1*4) + (10*5) = 72$$
$$t = (1*1) + (1*4) + (5*5) = 30$$
$$s = (5*1) + (1*2) + (1*5) = 12$$
$$\text{ostensive}_t = 28/72 = 0.417$$
$$\text{ostensive}_s = 12/72 = 0.167$$

Figure 1: Example ostensive calculation

3. EXPERIMENTAL SYSTEM

Our experiments used five systems. In this section we briefly outline the components that were common to all systems; in sections 5-7 we describe the specific variations of the experimental systems used in our experiments.

Our basic retrieval algorithm followed the approach given in (Ruthven, Lalmas and Van Rijsbergen, 2001). This assigns each term in the collection a set of weights. Each weight is calculated by a separate weighting scheme and reflects different aspects of how the term is used within the collection and individual documents. The retrieval score of a document is given by the sum of all the term weights of the query terms contained within the documents. This approach generally gives better results than the more standard *tf*idf* approaches (Ruthven et al., 2001).

After query expansion, sections 5-7, the RF systems traditionally weight query terms according to some measure of how useful they are in attracting relevant material (section 1). Our systems, instead, select which weighting schemes are best at indicating relevant material for each query term. This was shown to be preferable to assigning each query term a new weight based on relevance information (Ruthven et al., 2001).

In our system, searchers entered a natural language expression as a query and were shown the titles of the retrieved documents in groups of ten titles. A screen-shot of one of our interfaces is given in Appendix A, Figure A.1.

Clicking on the title displayed the full-text of the title with query terms highlighted in bold. The searchers were asked to mark any document that they felt contained useful information using the slider shown in Figure 2. We asked our subjects to assess the usefulness of documents, rather than the relevance, to encourage the subjects to make personal assessments on the relation between the documents and search tasks rather than make topical assessments of the match of the query and documents.

The relevance slider in Figure 2 was initially set to *Not useful* for each retrieved document. Unassessed documents were considered by default to be not useful to the searcher and counted as not relevant for the purposes of RF, (see section 5).

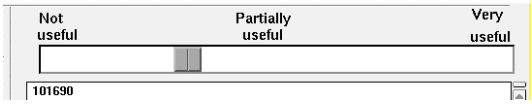

Figure 2: Relevance slider

4. EXPERIMENTAL DETAILS

4.1 Document collection

The document collection we used in our experiments consisted of a set of full-length newspaper articles, comprised of the LA Times and Financial Times collections from the TREC[4] initiative, (Voorhees and Harman, 2000). This gave a single collection consisting of over 340 000 documents.

4.2 Search tasks

The search tasks for these experiments were based on the topics used in the interactive track of TREC-6. We modified the topic descriptions, placing them within simulated situations as proposed by Borlund (Borlund, 2000). This technique asserts that searchers should be given search scenarios that reflect and promote a real information-seeking situation. The simulated situations allow a subjective and dynamic interpretation of relevance by the searcher. An example of one of the six simulated situations we used is given in Figure 3, the other five topics are given in Appendix B.

Several valuable paintings and other works of art in a local Glasgow museum have been discovered to be fakes. The museum's spokesman claims that art crime – in particular fraud – is becoming more common. He also claims that it is difficult to distinguish deliberate crime from genuine mistakes made by people selling works of art. You wonder if he is correct or whether these are excuses. You think more information on art crime, and on genuine cases of art fraud, can help you decide if the spokesman is correct.

Figure 3: Simulated situation

4.3 Experimental subjects

The subjects in our experiments were university students, 5 female and 13 male, with an average age of 23, and a variety of academic backgrounds. The subjects had experience of web search engines and library search facilities (average 4 years) but relatively little experience with any other IR system. No subject reported experience with IR systems that offered RF functionality.

4.4 Experimental methodology

Each experiment used two systems; a *control* system and an *experimental* system, discussed in sections 5-7. In each experiment six subjects each completed the same six search tasks; three tasks on the control system, three on the experimental system. The order of presentation of task and allocation of tasks to the control and experimental systems was randomised across the experimental subjects. No subject could take part in more than one experiment to limit familiarity with the search tasks and learning. In the experiments the subjects were given 15 minutes to search on each of the six search tasks.

The subjects were given a short tutorial on the main features of the system and were walked-through a sample search and then allowed to practice searching on the system. The subjects were instructed to search in any way they felt comfortable using the search systems and were encouraged to make their own assessments as to the utility of the documents found. The only specific task the subjects were given was to mark any useful document found. After each search the subjects were encouraged to discuss the search and the information they found whilst searching.

5. EXPERIMENT ONE: RF WITH F_4_*PO* AGAINST NO RF

Our first experiment compared the performance of RF incorporating our new term ranking algorithm, F_4_*po*, against no RF. In this experiment we were interested in how well the terms suggested by the system compared against the terms suggested by the searcher. In this experiment neither the control nor experimental system explicitly offered the subject a RF option; the subject was only offered a new search option. However, on the experimental system each time the subject performed a new search, the system implicitly performed a RF iteration[5]. That is, although the searcher asked for a new search, the system actually ran an iteration of RF instead. In the experimental system any new query terms added by the searcher were also included in the new search; the experimental hypothesis, then, was a comparison of searcher query modification (control system) versus searcher query modification plus RF query modification (experimental system).

The system therefore added terms to the query before retrieving a new set of documents but the changes to the query were hidden from the searcher. The query expansion method we used comes from (Ruthven, Lalmas and Van Rijsbergen, 2001). For each relevant document found the system adds the first expansion term, from the expansion term ranking, which appears in the document. This method of query expansion adds a variable number of terms to the query and was shown to be generally better than adding a fixed number of expansion terms to each query (Ruthven et al., 2001).

For the subject there was no observable difference between the two systems at the interface level: both systems appeared to do a new search each time. The only difference between the control and experimental system was the method by which the query was modified and the documents were ranked – the RF method of the experimental system. As we were interested only in the performance of RF against no RF the information regarding the initial search was excluded and the results from Experiment One only refer to the searches carried out after the initial search formulation for each search task. This allows a direct comparison of RF only against no RF.

	Topics					
Condition	**1**	**2**	**3**	**4**	**5**	**6**
Control	**31.67%**	4.07%	6.67%	4.07%	10.00%	13.33%
Experimental	7.78%	**6.00%**	**10.83%**	**11.67%**	**17.33%**	**20.83%**

Table 2: Results of documents relevant per retrieved
bold figures indicate higher value

The subjects carried out twice as many post-initial searches on the control than experimental system (2.28 per search task control, 1.56 experimental). Comparing the precision by measuring the number of documents assessed relevant by the number of documents retrieved, Table 2, it can be seen that the experimental system gives better precision for five of the six search topics.

A second comparison is to compare how many of the documents the subjects viewed were assessed as being relevant, shown in Table 3. Again, for the majority of topics the subjects found a higher proportion of relevant documents with the experimental (feedback) system. This was in spite of viewing the same proportion of retrieved documents on the control and experimental systems (12.94 documents viewed per search task on the control system, 13.67 documents on the experimental system).

Therefore the searchers are, generally, finding a higher percentage of relevant documents with the experimental system per documents retrieved and documents that the subject chooses to view. This shows that the F_4*_po* method

of ranking terms does work as a RF component: it does lead to better retrieval than the searcher's choice of query terms alone.

	Topics					
Condition	**1**	**2**	**3**	**4**	**5**	**6**
Control	**70.37%**	29.73%	34.78%	22.92%	**55.26%**	54.05%
Experimental	22.95%	**60.00%**	**56.52%**	**41.18%**	32.10%	**78.13%**

Table 3: Results of documents relevant per viewed
bold figures indicate higher value

6. EXPERIMENT TWO: F_4_*PO* AGAINST F_4 FOR RANKING CANDIDATE EXPANSION TERMS

In Experiment Two we compared the performance of the F_4_*po* method of ranking candidate expansion terms against the original F_4 term ranking technique. The intention is to see whether the F_4_*po* technique gives different results to those given by F_4. The interfaces for the control and experimental systems are identical and both offer an explicit RF option, Appendix A, Figure A.1. The retrieval and RF algorithms underlying the experimental interface for this experiment are the same as for Experiment One; the control RF system is identical to the experimental system except that it uses F_4 instead of F_4_*po*.

Details regarding the overall search behaviour of the subjects and the search effectiveness of the two systems are summarised in Table 4. All figures in Table are average values per search task.

	Control (F_4)	**Experimental (F_4_*po*)**
New search iterations	2.72	**2.89**
RF iterations	**2.00**	1.39
Documents viewed	**23.98**	19.67
Documents retrieved	**101.83**	97.17
Precision (relevant/viewed)	**54%**	49%
Relevant documents	**12.89**	9.56

Table 4: Overall search behaviour
bold figures indicate higher figures

The values given in Table 4 would appear to indicate a favour for the control system: the subjects performed more RF, viewed more documents and found more relevant documents per search task. However the subjects' perceptions of the terms suggested by the systems were in favour of the experimental (F_4_*po*) system. At the end of each search the subjects were asked how useful

the terms added by the system were to their search. This was on a 5-point scale, rated from 1 (*Not at all* (useful)) to 5 (*Extremely* (useful)). The average response when the subjects rated the terms suggested by the control system was 1.67 compared with 2.44 when the subjects used the experimental system. This value was found to be statistically significant (t =-2.80)[6]. That is, the subjects found the F_4*_po* terms more useful than the F_4*_standard* ones.

The subjects also informally, whilst searching, remarked on the more obvious nature of the F_4*_po* term suggestions. An example of the type of terms added by F_4 and F_4*_po* systems is shown in Figure 4. This example is drawn from a real search, chosen at random. The subject submitted the query '*hubble space telescope*' and marked four documents relevant at the first iteration. Figure 4 shows the top ten terms ranked by F_4 and F_4*_po*.

F_4	**F_4*_po***
accrete	astronomer
chaisson	hubble
cullers	telescope
goldreich	universe
sandpile	astronomers
terrile	telescopes
borucki	scientists
machtley	orbit
nebula	nasa
astronomer	earth

Figure 4: Example candidate expansion terms ranked by F_4 and F_4*_po*

The F_4 algorithm selected terms that are less usual in the collection (*accrete*, *chaisson*) whereas the F_4*_po* algorithm selected variants of existing terms (*telescopes*), and more obvious terms (*orbit*, *nasa*, *earth*). The F_4*_po* algorithm also returned the original query terms higher up than the F_4 algorithm.

A further analysis was used to uncover how the expansion terms were actually treated by the subject: were the expansion terms often retained or removed by the subject. One justification for this kind of analysis is that subjects may be put off using RF because the suggested terms do not appear useful. Consequently they may lose out on the potential benefits from RF. The results of this analysis are summarised in Table 5.

In Table 5 (rows 3 and 4) we show the source of query terms that were added after the initial query: either added by the subjects (row 3) or the system through RF (row 4). We also show how many of the terms the subjects added were removed later by the subjects (row 5) and how many terms added by the

system were removed by the subjects (row 6). The figures are averaged over search tasks.

	F_4	**F_4_*po***
Source of added terms		
subject	2.00	**2.33**
system	**3.33**	1.11
Source of removed terms		
subject	0.72	**1.17**
system	**2.28**	0.67

Table 5: Summary of query term addition and removal per search task
bold figures indicate higher value

Comparing the two systems, Table 5 shows that the subjects added more of their own terms per task with the experimental system and RF added more terms when using the F_4 than the F_4_*po* algorithm per feedback iteration The main reason for the latter finding is that the F_4_*po* function in the experimental system prioritises the original query terms more than the F_4 algorithm, and is likely to add fewer expansion terms to the query. This also, perhaps, relates to the fewer RF iterations performed on the experimental system. A different query expansion algorithm should be tested here to elicit any relation between the number of terms added and the number of RF iterations performed as it may be the case that the subjects were performing less RF as RF was making fewer query term changes in the experimental system.

The difference between the number of the subjects' own terms removed was not significant (t = -1.16). However the difference between the number of *system* suggested terms removed was significant (t = 2.54). This latter finding suggests that the terms suggested by the F_4_*po* system were felt to be better search terms by the subject. Although the F_4_*po* system did not improve more queries or give better overall results it was seen by the subjects as a better term suggestion technique. The next experiment tests the effectiveness of the two term ranking schemes when the subject is selecting new query terms – Interactive Query Expansion.

7. EXPERIMENT THREE: F_4_*PO* AGAINST F_4 FOR INTERACTIVE QUERY EXPANSION

The third experiment compared the effectiveness of the F_4 and F_4_*po* term ranking schemes in suggesting new expansion terms for selection by the subjects. In this experiment the control system used the F_4 algorithm to suggest 20 possible expansion terms and the experimental system used the F_4_*po* algorithm to suggest expansion terms. Both control and experimental

systems used the same interface, the only difference between the two systems was the underlying term suggestion technique. The interface is shown in Appendix A, Figure A.2.

In this experiment we are primarily interested in how the subjects used the suggested expansion terms compared with how they used their own terms. In Table 6 we present details on how the subjects added or removed query terms based on the source of the query term.

From Table 6, there were differences in how the subjects added new query terms. For example, in the control system the subjects were more likely to add their own terms to their query than ones suggested by the system (8.83 own terms added vs 1.61 system suggested expansion terms). On the experimental system, however, this was reversed: the subjects were more likely to add terms suggested by the system (6.67 own terms added vs 8.17 system suggested expansion terms).

The difference between the number of their *own* terms the subjects added was not significant ($t = 0.69$) however the difference in the number of the *system*-suggested terms added was significant ($t = -3.16$). That is, subjects were more likely to use the system-suggested terms when the system used the F_4*_po* term algorithm to suggest terms.

	Topics						
Control	**1**	**2**	**3**	**4**	**5**	**6**	**Averages**
Own terms added	26	8	**26**	**20**	**64**	15	**8.83**
System suggested term added	4	2	9	4	4	6	1.61
Own terms removed	16	**6**	**29**	**18**	**63**	0	**7.33**
System suggested term removed	1	**2**	**9**	**1**	2	0	**0.83**
Experimental	**1**	**2**	**3**	**4**	**5**	**6**	**Averages**
Own terms added	**31**	**14**	**26**	16	11	**22**	6.67
System suggested term added	**36**	**12**	2	**29**	**33**	**35**	**8.17**
Own terms removed	**20**	4	23	2	10	**10**	3.83
System suggested term removed	**2**	0	2	0	**6**	**0**	0.56

Table 6: Statistics on query terms in Experiment Three
bold figures indicate higher value

Next, we investigate whether the increase in term use led to an increase in retrieval effectiveness. In Table 7 we present the number of unique relevant documents found on average per topic and the average relevance score given

by the subjects to the documents they assessed as relevant. From Table 7, it can be seen that on all topics, with the exception of topic 3, the subjects found at least as many relevant documents on average and the average relevance score given to the documents found was higher. The difference between numbers of documents found was not significant (t = -0.69). However the difference between the average score given to a relevant document was significant (t = -5.29). These results indicate that, although the F_4*_po* suggested terms did not help find significantly more relevant documents, the F_4*_po* terms helped find *better* relevant documents[7].

		Topics						
		1	**2**	**3**	**4**	**5**	**6**	**Avg**
Control	Relevant documents	10.00	**8.00**	**12.33**	7.33	9.67	8.00	9.22
Exptl	Relevant documents	**11.00**	**8.00**	7.00	**9.33**	**21.67**	**9.33**	**11.06**
Control	Average relevance score	3.78	5.37	5.14	5.05	4.49	4.31	4.69
Exptl	Average relevance score	**6.91**	**6.82**	**6.01**	**7.33**	**7.08**	**5.48**	**6.61**

Table 7: Comparison of relevant documents found and average relevance score
bold figures indicate higher value. Exptl = experimental system, Avg = average

This also accords with the subjects' perceptions of the suggested expansion terms. As in Experiment Two we asked the subjects, after each search, how useful they thought were the terms suggested by the systems. As seen in Table 8 where the average response per topic for this question is shown, the subjects reported the terms suggested by the experimental, F_4*_po,* system as being more useful than the control, F_4, suggested terms. This difference held across the search tasks and the difference is statistically significant (t = -3.73).

	Topics					
Utility of terms	**1**	**2**	**3**	**4**	**5**	**6**
Control	1.33	2.33	1.33	1.67	2.00	2.00
Experimental	**3.33**	**2.67**	**1.67**	**3.67**	**4.50**	**5.00**

Table 8: Comparison of subject responses in Experiment Two regarding term utility
bold figures indicate higher value

This experiment showed that the terms suggested by the F_4*_po* weighting scheme could give better term suggestions: those that were preferred by the subjects and which lead to the retrieval of better relevant documents.

8. DISCUSSION AND CONCLUSIONS

The previous experiment showed the subjects used more expansion terms that were suggested by the F_4*_po* function. However, as noted throughout the experiments the use of the F_4*_po* function did not necessarily increase the retrieval of more relevant documents over the F_4*_standard* function. That is, whether used interactively or automatically the terms chosen by the two functions perform in a similar fashion. This is the case even though the terms ranked highly by the two algorithms are often very different.

To demonstrate this we took, for each of the subjects' searches, all the documents marked relevant by the subject and used these documents to create two lists of expansion terms; one list ranked by F_4*_standard* and one ranked by F_4*_po*. We then compared the top 20 terms in each list – the ones presented in interactive query expansion – and compared the overlap between the two lists, i.e. how many terms appeared in both lists. The results, Table 9, are averaged over all searches on a topic and show that, for an individual search, the two term ranking algorithms will only share around three terms (column 8). This means that the terms at the top of the expansion term ranking – the ones most likely to be used in query expansion – are different. Therefore even though different terms are being added to the query, similar retrieval results are being obtained. This is an issue that requires further investigation. In particular we should consider the searchers' intention behind selecting individual terms and the effect the searcher intends on the kind of documents being retrieved.

	Topics						
	1	**2**	**3**	**4**	**5**	**6**	**All topics**
%age overlap	19.83%	14.33%	11.17%	17.17%	16.83%	6.67%	14.33%
Shared terms	3.97	2.87	2.23	3.43	3.37	1.33	2.87

Table 9: Overlap between top 20 terms suggested by F_4*_standard* and F_4*_po* functions

Table 9 demonstrates that the terms suggested by the two term ranking algorithms are different. The subjects' perception of the two term ranking techniques, section 7, show that the subjects also perceive a difference regarding the terms' utility. However, the subjects' perceptions regarding a term's utility for searching do not necessarily match their judgements on the relevance of documents containing the terms, and neither does the subjects' perceptions on the search effectiveness of the systems used. In Experiments Two and Three we asked the subjects to assess their satisfaction on their

search[8]. The results for the systems that used F_4_*standard* were both lower than for those systems that used F_4_*po* (Experiment Two 3.05 F_4_*standard* vs 3.44 F_4_*po*, Experiment Three 2.72 F_4_*standard* vs 3.83 F_4_*po*). Therefore the main strength of the F_4_*po* function is that the terms it suggests are preferred by searchers.

The main contribution in this paper was a new method of ranking candidate expansion terms based on relevance information that incorporates partial relevance assessments and ostensive information. There are limitations to our experiments. In particular we used a small number of searchers, and a limited set of search tasks. These both limit the conclusions we can draw from these experiments. In addition we only used one set of interfaces. The presentation of documents, the method by which searchers assess documents and how expansion terms are presented to the searcher are obviously important factors in the *use* of RF and query expansion techniques. The presentation of interactive query expansion, for example, has been shown to be an important variable in the success and uptake of query expansion in (Koenemann and Belkin, 1996) and (Beaulieu, 1997).

Finally, we only attempted to incorporate behavioural information into one term ranking algorithm. We cannot guarantee that similar results will be obtained from other algorithms without further investigation. These set of experiments are intended to be viewed as a proof-of-concept investigation to investigate the general principle of incorporating user search information into the term ranking principle.

Our experiments indicate that our term ranking function performs well in RF, selects terms that are preferred by the searcher in automatic query expansion, and suggests better terms for interactive query expansion. This shows that incorporating information on the user's search activity *can* improve RF algorithms but we do require much more investigation to provide robust methods of connecting the search to the system. We hope that this initial investigation will promote interest in this area.

NOTES

[1] Usually all the terms which appear in at least one relevant document.
[2] F4_p(artial)o(stensive)
[3] The maximum relevance score that can be assigned to a document is 10. Therefore n_t becomes the maximum total relevance score that can be assigned to the set of documents containing term *t*
[4] The interactive TREC track is an initiative intended to investigate search systems with a highly interactive nature. More information is available at http://trec.nist.gov/

[5] Not including the initial search.
[6] Measured using a *t*-test for related samples, $p < 0.05$
[7] These figures and the ones regarding term utility in Table 8 are only for searches in which the subject used the term suggestion option. Out of the 18 searches on each system, 3 searches on the control system and 2 searches on the experimental system did not include use of the term suggestion option. All subjects used this option in the majority of their searches.
[8] The assessment was a score from 1 – 5 with 5 reflecting the highest satisfaction with the search.

ACKNOWLEDGEMENTS

We would like to thank Mark Dunlop and Pia Borlund for their helpful comments on the experimental design. We are especially grateful for the many useful comments from the anonymous referees which helped improve the content of this paper. This work was completed as part of the Library and Information Commission project Retrieval Through Explanation http://www.dcs.gla.ac.uk/ir/explanation, whilst the first author was at the University of Glasgow.

REFERENCES

Beaulieu, M. *Experiments with interfaces to support query expansion*. Journal of Documentation. **53**. 1. pp 8-19. 1997.

Borlund, P. *Experimental components for the evaluation of interactive information retrieval systems*. Journal of Documentation. **56**. 1. 2000.

Campbell, I. and van Rijsbergen, C. J. *The ostensive model for developing information needs.* In Proceedings of 2nd International conference on Conception of Library and Information Science (COLIS2). pp 251-268. 1996.

Efthimiadis, E.N. *User choices: A new yardstick for the evaluation of ranking algorithms for interactive query expansion*. Information Processing and Management, **31**. 4. pp 605-620. 1995.

Florance, V. and Marchionini, G. *Information processing in the context of medical care.* Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle. pp 158-163. 1995

Harman, D. *Relevance feedback revisited*. Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen. pp 1-10. 1992.

Heuer, R. J. Jr. *Psychology of intelligence analysis*. Center for the Study of Intelligence. Central Intelligence Agency. 1999

Koenemann, J. and Belkin, N. J. *A case for interaction: a study of interactive information retrieval behavior and effectiveness*. Proceedings of the Human Factors in Computing Systems Conference (CHI'96). pp 205-212. Zurich. 1996.

Robertson, S. E. and Sparck Jones, K. *Relevance weighting of search terms*. Journal of the American Society for Information Science. **27**. pp. pp129-146. 1976.

Ruthven, I., Lalmas, M., and van Rijsbergen. C. J. *Empirical investigations on query modification using abductive explanations*. Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans. 2001.

Salton, G. and Buckley, C. *Improving retrieval performance by relevance feedback*. Journal of the American Society for Information Science. **41**. 4. pp 288-297. 1990.

Spink, A., Greisdorf, H., and Bateman, J. *From highly relevant to not relevant: examining different regions of relevance.* Information Processing and Management. **34.** 5. pp 599-621. 1998.

Tiamiyu, M. A. and Ajiferuke, I. Y. *A total relevance and document interaction effects model for the evaluation of information retrieval processes*. Information Processing and Management. **24**. 4. pp 391-404. 1988.

Vakkari, P. *Cognition and changes of search terms and tactics during task performance: a longitudinal study*. Proceedings of the RIAO'2000 Conference. Paris. pp 894-907. 2000a.

Vakkari, P. *Relevance and contributing information types of searched documents in task performance*. Proceedings of the twenty-third annual international ACM SIGIR Conference on Research and development in information retrieval. pp 2-9. Athens. 2000b.

Voorhees, E. and Harman, D. *Proceedings of the sixth Text REtrieval Conference (TREC-6).* National Institute of Standards and Technology (NIST), Special Publication 500-240. 1997.

Appendix A

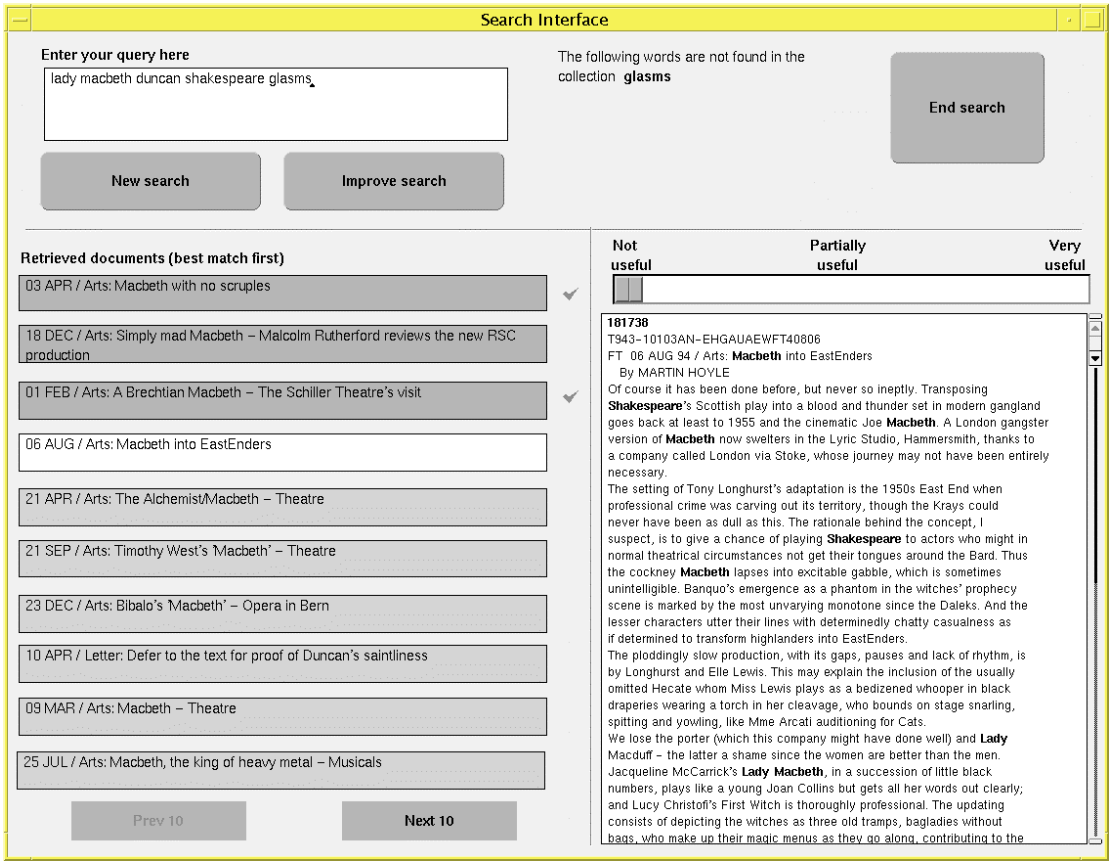

Figure A.1: Interface for Experiment Two

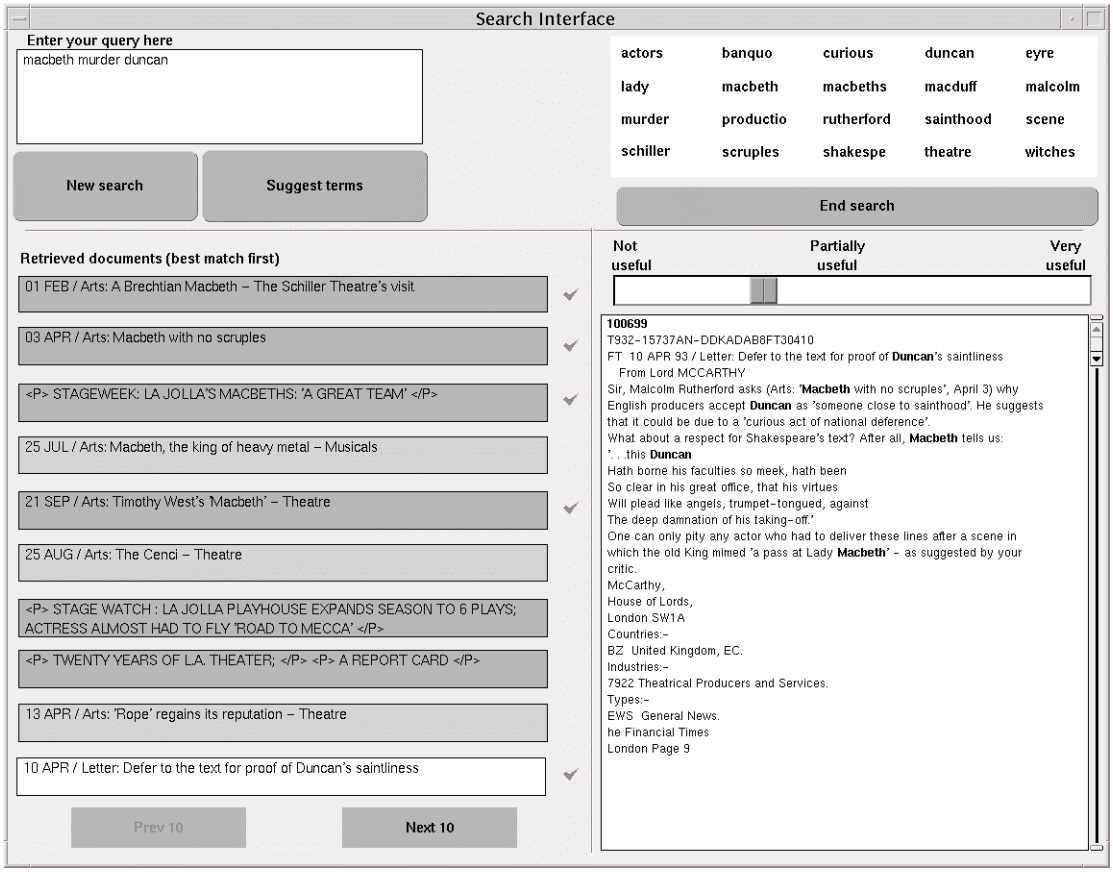

Figure A.2: Interface for Experiment Three

Appendix B

Topic one

At a recent party you overhear a discussion about whether science funding gives value for money. One person claimed that many expensive projects, such as the Hubble Telescope, do not produce significant positive advances. You are not sure how true this statement is, and would like to find more information on the positive achievements of the Hubble Telescope since it was launched in 1991.

Topic two

The new Scottish Parliament is considering planning permission for a series of large hydroelectric projects. These projects will use water power to produce electricity for a large area of Scotland. Supporters of the projects claim that they will give cheaper electricity and reduce global-warming, opponents argue that the projects may cause environmental damage and harm tourism. The Parliament has decided to hold a vote for all Scottish residents to decide if

these projects should go ahead. You have little independent information upon which to base your decision, and would like information on similar projects.

Topic three
It is likely that a British General Election will be held in May this year. In the last General Election, one of the main issues was the relatively low number of female members of parliament. This prompted one party to introduce special measures to increase the number of female candidates in the election. Other politicians argue that poor representation of women in parliament is not a specific feature of British politics. As the poor representation is likely to be a major issue in the forthcoming election, you would like to be more informed about the representation of women in politics.

Topic five
You and a friend are trying to choose a holiday for later this summer. One possible holiday destination will mean taking several ferry trips but you have heard rumours that ferries in this area have a poor safety record. You need to book your holiday soon but need more information on the dangers of ferry travel.

Topic six
Your best friend is an active member of a major wildlife preservation group. She is working on a project to build an electronic database of wildlife species that are in danger of extinction and the steps that different countries have taken to protect these species. She has asked you for help in providing information on international attempts to save native species, and the causes of wildlife extinction.