Strathprints Institutional Repository

Ruthven, I. and Lalmas, M. and van Rijsbergen, C.J. (2003) *Incorporating user search behaviour into relevance feedback.* Journal of the American Society for Information Science and Technology, 54 (6). pp. 528-548. ISSN 15322882

Incorporating user search behaviour into relevance feedback

Ian Ruthven[*1], Mounia Lalmas[2] and Keith van Rijsbergen[3]
[1]Department of Computer and Information Sciences, University of Strathclyde, Glasgow, G1 IXH
[2]Department of Computer Science, Queen Mary, University of London, London, E1 4NS
[3]Department of Computing Science, University of Glasgow, Glasgow, G12 8QQ

Ian.Ruthven@cis.strath.ac.uk, mounia@dcs.qmul.ac.uk, keith@dcs.gla.ac.uk

Abstract

In this paper we present five user experiments on incorporating behavioural information into the relevance feedback process. In particular we concentrate on ranking terms for query expansion and selecting new terms to add to the user's query. Our experiments are an attempt to widen the evidence used for relevance feedback from simply the relevant documents to include information on how users are searching. We show that this information can lead to more successful relevance feedback techniques. We also show that the presentation of relevance feedback to the user is important in the success of relevance feedback.

1 Introduction

The majority of Information Retrieval (IR) systems require users to enter a query to initiate a search. However, queries are often *imprecise* representations of the user's information. Relevance feedback (RF) techniques aim to improve a user's query by using documents that have been assessed relevant by the user (Harman, 1992). RF is generally composed of three stages; the system first selects possible expansion terms 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*). The possible expansion terms themselves come from the assessed relevant documents.

However the relevant *documents* themselves only form part of the relevance information given by the user whilst searching. For example, the stage in a search when a user marks a document relevant can give information on what the user *currently* finds relevant (Campbell & Van Rijsbergen, 1996), and the relevance score a user gives to a document can give information on *how* relevant the document is to the user, (Spink, Greisdorf & Bateman, 1998). Other potentially relevant information comes from the search iteration as a whole, e.g. how many documents have been assessed relevant, where in the document ranking the relevant documents have been found, and how similar are the relevant documents. This kind of *behavioural* information can act as an important source of evidence on what the user finds to be relevant.

In this paper we present five user experiments on incorporating behavioural information into the RF process. In particular we concentrate on selecting new terms to add to the user's query (*query expansion*) and ranking terms for query expansion (*term ranking*). In section 3 we show how we

[*] Corresponding author. This work was completed whilst the first author was at the University of Glasgow.

incorporate behavioural evidence into the term ranking process and in section 8 we deal with query expansion. In each section we shall present the motivation for the investigation. In sections 4-6 and sections 9-10 we describe the experiments associated with each investigation. Prior to this, in section 2, we present the overall experimental environment that was used in the experiments as these are common to both sets of experiments. We conclude in section 11 with a discussion of the overall research.

2 Experimental details

In this section, we outline the components of our experiments. In section 2.1 we discuss the document collection that we used, in section 2.2 we discuss the search tasks and in section 2.3 we discuss the experimental procedure that was followed in the experiments and the experimental subjects. In section 2.4 we give a general discussion on how the results will be analysed, and in section 2.5 we give an overview of the experiments.

2.1 Document collection

For the experiments reported in this paper we used the Financial Times (**FT**) and Los Angeles Times (**LA**) collections from the TREC initiative (Voorhees & Harman, 2000). The FT collection consists of full-length newspaper articles from the Financial Times of London published from 1991 – 1994. The LA collection consists of a sample of approximately 40% of the articles published by this newspaper in the period from January 1989 to December 1990. The combined collections gives a document set of over 340 000 documents (Table 1, column 4), which covers the period 1989 – 1994. This cannot be regarded as a set of *currently* topical documents and subjects would not be able to search using current new events. However, the collection is not out-of-date as regards the search situations given to the subjects (section 2.2).

	FT	**LA**	**Combined**
Number of documents	210 158	131 896	342 054
Average document length (index terms)	412	526	456
Number of unique terms in the collection	245 678	244 874	375 295

Table 1: Document collections

2.2 Search tasks

The search tasks given to our experimental subjects were based on search topics taken from the interactive track of TREC-6[1] (Over, 1998). We chose these topics as they have previously been used with our document collections and the modifications we made upon the topics have been investigated elsewhere (Borlund & Ingwersen, 1999).

INTTREC6 used six topics for the interactive task. We retained five of these topics (topics numbered[2] 303i, 307i, 326i, 322i, 347i[3]). Topic 339i, which asked subjects to search for information on '*Alzheimer's drug treatment'*, was excluded. This decision was made based on previous use of these topics by Borlund and Ingwersen whose experience suggested some

[1] Hereafter shortened to INTTREC6 for convenience.

[2] The topic numbers relate to the TREC-6 non-interactive *ad-hoc* track, which uses fifty topics (Voorhees and Harman, 2000). The INTTREC6 track selected a number of these for interactive searching.

[3] The topic titles are 'Hubble telescope achievements', 'New hydroelectric projects', 'Women in parliament', 'International art crime', 'Ferry sinkings', and 'Wildlife extinctions'.

searchers may feel uncomfortable searching on this topic (Borlund & Ingwersen, 1999). This topic was replaced by the TREC-6 ad-hoc topic number 321, '*Women in Parliaments'*. This topic was chosen as it avoided any potentially distressing themes, was not similar to any of the existing topics and was not a topic that required specialist knowledge of a domain. The INTTREC6 search topics were placed within *simulated work-task situations* as proposed in (Borlund & Ingwersen, 1999; Borlund, 2000, 2000b). This technique, developed by Borlund, asserts that experimental subjects should be given search scenarios that reflect and promote a real information-seeking situation (Borlund, 2000).

The simulated situations, such as the one shown in Figure 1 are intended to achieve two main objectives (Borlund, 2000b). First, they are aimed at promoting a simulated information need in a subject. That is, the simulated situation should engage the subjects in the search by the identification of the subject with the situation. Second, the simulated situations position the search within a realistic context. This allows the experimental subject to provide his or her own interpretation of what information is required and allows the subject to develop the information need naturally.

The use of simulated situations therefore encourages more realistic searching on behalf of the experimental subjects whilst retaining experimental control.

> *Several valuable paintings and other works of art in a local Glasgow museum have been discovered to be fakes*[4]*. 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 1: Simulated situation for INTTREC6 topic 322i[5]

2.3 Experimental methodology

In this section we describe the experimental procedure we followed for our experiments. The same methodology was used for each of the five experiments, the only difference being the systems used in each experiment, and the subjects used in each experiment[6].

Each subject was asked to perform a search on each of the simulated situations, performing three simulated situations on a *control* system and three on an *experimental* system. The control and experimental systems were different for each experiment as will be explained in sections 4-10.

Each subject was given a maximum of 15 minutes to search on each task. The order in which the situations were presented, and the choice of which system a subject used for each search, was determined by an experimental matrix. The matrix used in the experiments described here, Figure 2, permutates order of situations, distribution of situations across systems and order of systems.

[4] This situation is not based on a real event.

[5] We retain the original TREC topic numbers to differentiate the simulated situations.

[6] No subject could take part in more than one experiment. This was to avoid the subjects becoming familiar with the search tasks and to control learning on behalf of the subject regarding the systems used.

The experiments described in this paper used six experimental subjects per experiment. This number of subjects does not allow a complete randomisation of subject, system and simulated situation so we have concentrated on randomisation of order in which subjects were presented the simulated situations and system. The same matrix was used for all experiments.

Subject	**Situation**	**Situation**	**Situation**	**Situation**	**Situation**	**Situation**
1	**303i**	*321*	**326i**	*307i*	**322i**	*347i*
2	*307i*	**322i**	*347i*	**321**	*326i*	**303i**
3	**307i**	*347i*	**326i**	*321*	**303i**	*322i*
4	*322i*	**307i**	*321*	**347i**	*303i*	**326i**
5	*326i*	**321**	*303i*	**322i**	*307i*	**347i**
6	**347i**	*322i*	**307i**	*326i*	**321**	*303i*

Figure 2: Experimental matrix
where **bold** figures = simulated situations to be run on the experimental system,
italic figures = simulated situations to be run on the control system

In each experiment the subject was given a tutorial on the search systems, were allowed to practice searching on the system, then were presented the simulated situations for searching. The subjects were asked to imagine they were the person described in the simulated situation and asked to find information that they thought would be useful for the simulated situation. To encourage the subjects to search in a naturalistic way, we did not ask our subjects to view all documents retrieved or read the whole text of any documents that they chose to examine. Rather we asked them to search in any way they felt comfortable.

Each experiment used a different interface, each of which were based on the simple interface shown in Figure A.1, Appendix A. The differences between the interfaces are discussed within the context of each experiment, sections 4-10.

In all the systems used in our experiments, users entered natural language expressions as queries and were shown the titles of the retrieved documents in groups of ten titles. Clicking on a title displayed the full-text of the corresponding document with any query terms contained in the document highlighted in bold. The users were asked to mark any document that they felt contained useful information using the slider shown in Figure 3. In our experiments we asked our subjects to assess the utility of documents, rather than the relevance, to encourage the subjects to make personal assessments on the relation between the documents and the search tasks.

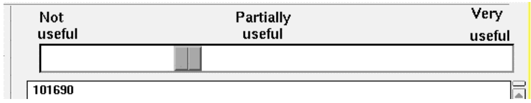


Figure 3: Relevance slider

The experimental subjects themselves were students in the Computing Science Department at the University of Glasgow. Half of the subjects were undergraduate computing students, and half were students on the Masters in Information Technology course. These latter students had previous degrees in a non-computing discipline. Thirty students took part in the experiments; 9 of the subjects were female, 21 male, and their average age was 23.

The subjects had relatively high experience of on-line searching (average 4.28 years), which was mostly gained through library search facilities and web search engines. The subjects reported good experience on these two forms of IR system but little experience of any other search system such as conventional text retrieval systems. The subjects were also relatively frequent searchers searching daily or at least weekly. All had good previous experience of point-and-click interfaces such as the ones used in these experiments. No subject reported experience of an IR system that offered RF functionality.

2.4 Analysis

For each experiment we analyse the results under three main headings. The first examines the subjects' overall search *behaviour*; this analysis looks for changes in how subjects searched on the control and experimental system. The second examines the search *effectiveness* of the two systems: did the subjects have a more effective search on the control or experimental system? Finally we examine the subjects' perceptions of the two systems: did the subjects prefer one system over the other in each experiment? Where appropriate we also examine differences before and after feedback to isolate the effect of the feedback techniques on the search. Tests for statistical significance will be used for important results. Specifically we use a paired *t*-test for related samples, comparing subject aggregate performance on each situation using the control and experimental system.

2.5 Overview of experiments

In this paper we present five experiments, in two sets. The first set of experiments – Experiments One, Two and Three – look at incorporating user behaviour into the process of ranking terms for query expansion. Specifically we propose a method for incorporating information on the partial relevance of documents and the temporal relevance of documents into the process of deciding which terms may be useful for query expansion.

We compare this method of term ranking against no RF (experiment one), and against a standard method of ranking expansion terms for both automatic RF (experiment two) and interactive RF (experiment three).

The first set of experiments investigates how terms should be ordered for query expansion; the second set of experiments investigates how terms should be *selected* for query expansion. In the second set of experiments we investigate a technique for using search behaviour to select how expansion terms should be selected. This method involves selecting, from a number of possible methods of query expansion, which method is most appropriate for an individual search. Experiment Four compares this method of selecting expansion algorithms against a non-selective query expansion technique. The final experiment, experiment five, investigates the effect of giving the user more information on how the selection procedure operates. Experiment Five is intended to investigate the role of searcher knowledge of RF operations on the use of RF by searchers.

3 Term ranking and user behaviour

In this section we concentrate on the first stage of RF – 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 terms 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 user can easily edit the reformulated query manually.

However it is important to use a good method of ranking terms to reflect their possible utility in retrieving relevant information. One of the main claims of this paper is that behavioural information – information on how users make relevance assessments – can help improve RF. In this section we show how behavioural evidence can be incorporated within the term ranking process. In particular we investigate the role of *ostension* –varying the importance of documents according to when they were assessed relevant - and the use of *partial* relevance assessments.

In the experiments described in this paper we investigate this by developing an extension to the standard F_4 term ranking[7] algorithm (Robertson & Sparck Jones, 1976), section 3.1. The extension to F_4 will be called F_4*_po*[8] and the original F_4 algorithm will be referred to as F_4*_standard*.

The F_4*_po* algorithm incorporates information from two sources: partial relevance assessments, information on the degree of relevance of a document to a search (Spink, Greisdorf & Bateman, 1998), and ostensive evidence, information on *when* in a search the searcher regarded documents as relevant (Campbell & Van Rijsbergen, 1996). The weight of a term is composed of two components, one of which calculates the contribution coming from the partial evidence and one that reflects the contribution coming from the ostensive evidence. The weight of a term comes from the product of these two components, as shown in Equation 1 for term *i*.

$$F_4\_po_i = partial_i * ostensive_i$$

Equation 1: F_4*_po* term ranking scheme

In section 3.1 we describe how the *partial$_i$* component is implemented and in section 3.2 we describe how the *ostensive$_i$* component is implemented.

3.1 Incorporating partial relevance assessments - *partial$_i$* component

The *partial$_i$* component is based on the F_4*_standard* scheme. The F_4*_standard* term ranking scheme, Equation 2, treats relevance as a binary decision, i.e. all relevance assessments were taken to have a value of 1 (relevant) or 0 (non-relevant).

$$w_i = \log \frac{r_i / (R - r_i)}{(n_i - r_i) / (N - n_i - R + r_i)}$$

Equation 2: F_4*_standard* term ranking scheme

[7] The F_4 algorithm was designed to weight terms by the use of relevance information, i.e. it was used as a term *weighting* function. However it is often used to rank terms for query expansion, e.g. (Efthimiadis, 1995), i.e. used as a term *ranking* function. As the main interest in this paper is to investigate how terms should be ranked for query expansion we refer to this function as a term ranking algorithm.

[8] F_4 *_p(artial)o(ostensive)*

In all the experiments described in this paper the subjects were asked to mark, using the slider shown in Figure 3, *how* useful a document was to their search. Internally the position of the slider corresponds to a number on a scale of 0-10[9]. These non-binary assessments are incorporated into the F_4 term ranking scheme by treating the value assigned to the document as *part* of a relevance assessment. A document that received a value of 10 was treated as a complete relevant document, a document that received a value of 5 was treated as half a relevant document, a document that received a value of 1 was treated as a tenth of a relevant document, and so on. The aim is to test whether partial assessments can give better estimates of term utility than binary assessments.

Table 2 outlines the conversion from the binary, F4 *_standard* weight to the partial, F_4*_po*, weight.

	F4 *_standard*	**F4 *_po***
r_i	number of relevant documents containing term *i*	sum of relevance assessments of documents containing term *i*
R	number of relevant documents	sum of relevance assessments given in search
n_i	number of documents containing term *i*	number of documents containing term *i* multiplied by maximum relevance assessment
N	number of documents in collection	number of documents in collection multiplied by maximum relevance assessment

Table 2: Conversion from binary F_4*_standard* to partial F_4*_po*

Table 3 gives examples of the difference between F_4*_standard* and F_4*_po*. This example is based on calculating the weight for term *i* which appears in 10 documents, 3 of which have been assessed relevant. The collection contains 100 documents, 7 of which have been assessed relevant. The relevance scores for the relevant documents containing term *i* are shown in the column labelled *rel_i*; the relevance scores for the relevant documents that do *not* contain term *i* are assumed to be 1 for the sake of simplicity.

	rel_i	**r_i**	**n_i**	**R**	**N**	**Weight**
F4 _standard	1,1,1	3	10	7	100	2.22
F_4*_po*	1,1,1	3	100	7	1000	1.94
F_4*_po*	3,5,7	15	100	19	1000	3.68
F_4*_po*	10,10,10	30	100	34	1000	4.56

Table 3: Example comparison of binary F_4*_standard* to partial F_4*_po*

Rows 2 and 3 compare the effect of F_4*_standard* and F_4*_po* on the same set of relevance assessments. Both term ranking algorithms give positive weights to term *i*. If we vary the relevance scores given to the relevant documents containing term *i,* as in rows 4 and 5, the weights assigned to term *i* change. Specifically the weights change according to how much the

[9] 0 was the default value indicating not relevant, values of 1-10 were taken to indicate relevant, or useful, material.

documents containing *i* contribute to the *overall* relevance assessments given by the user. For example, in row 5, the user has found three documents containing term *i*. Each of these documents has been given the maximum relevance score of 10. The remaining four relevant documents (the ones that do not contain *i*) have only been given a relevance score of 1. Therefore out of a total relevance assessment of 34 (10*3 + 1*4), 30 units of assessment have come from the documents containing term *i*. Term *i*, accordingly, is assessed to be very useful term and given a high score. If, however, the documents containing term *i* contribute less to the overall relevance assessments, as in row 4, term *i* receives a lower weight.

Therefore varying the way F_4*_standard* interprets the relevance information given by partial relevance assessments can lead to a more detailed method of ranking terms.

3.2 Incorporating partial relevance assessments – *ostensive$_i$* component

Campbell and Van Rijsbergen argue that *when* in a search a document was marked relevant should be treated as important (Campbell & Van Rijsbergen, 1996). This allows for the fact that a user may change his or her criteria for relevance when encountering newly retrieved material. Therefore the documents most recently marked relevant are more indicative of what the user currently finds relevant – provide more *ostensive* evidence as to relevance. Campbell investigated an application of using ostensive evidence for image retrieval (Campbell, 1990). In Campbell's experiments users searched on relatively static information needs; in this paper we investigate the ostensive evidence for dynamic search situations.

In the experiments, although the subjects had a limited time to perform each search (15 minutes, section 2.3), they could run as many searches or feedback iterations as they felt necessary. This allowed us to investigate the potential effect of ostensive evidence: weighting terms according to *when* users indicated relevant material. Ostensive evidence was incorporated into the term ranking algorithm by a similar means to the partial evidence. The equation used to calculate the ostensive value of the term is shown in Equation 3.

In Equation 3 the ostensive weight of term *i*, is based on a proportion of the ostensive evidence for term *i* 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 *i*. The ostensive evidence for term *i* is the sum of the relevant documents containing term *i* multiplied by the iteration in which the documents were marked relevant. Therefore the more relevant documents term *i* appears in, the higher weight it receives and the more recently-viewed relevant documents term *i* appears in the higher weight it receives.

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

Equation 3: Calculation of ostensive weight
where s = total number of feedback iterations, r_{ji} = number of relevant documents containing term i in iteration j, $\max_{ostensive}$ = maximum possible ostensive evidence

An example of this is shown in Figure 4, for two terms – term *t* and term *q*, based on the data given in Table 4. In Table 4, 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*, (**r**$_t$, row 3), and some of which contain term *q* (**r**$_q$, row 4).

	Iterations of feedback					
	1	**2**	**3**	**4**	**5**	**Total**
r$_t$	1	0	0	1	5	7
r$_q$	5	1	0	0	1	7
R	5	2	3	1	10	21

Table 4: Example ostensive data

$$\text{max\_ostensive} = (5*1) + (2*2) + (3*3) + (1*4) + (10*5) = 72$$

$$t = (1*1) + (1*4) + (5*5) = 30$$

$$q = (5*1) + (1*2) + (1*5) = 12$$

$$\text{ostensive}_t = 28/72 = 0.417$$

$$\text{ostensive}_q = 12/72 = 0.167$$

Figure 4: Example ostensive calculation

The value of max$_{ostensive}$ is identical for both terms: both terms could have appeared in all the relevant documents at all iterations. The incorporation of the ostensive evidence allows the F4*_po* algorithm to incorporate when the documents containing term *t* or *q* were marked relevant. Even though both terms appear in the same number of relevant documents, term *t* receives a higher score as it appears in more of the documents that were marked relevant in the recent search iterations.

The partial component of the F4*_po* weight is multiplied by the ostensive weight to give a final weight for each term. Terms are then ranked in decreasing order of this weight to reflect how useful they are at discriminating the user-selected relevant documents. Terms that are given high F4*_po* weights are those that appear in more of the documents the searcher has recently marked as being highly relevant; those terms that receive low F4*_po* weights are those that appear in fewer, less relevant, and less recent documents.

The new weighting scheme will be investigated in several experiments, described in the following sections. For convenience of exposition, in each experiment we label one system as the *control* system and one system as the *experimental* system.

4 Experiment One

The first experiment investigated the performance of RF using the F4*_po* term ranking technique against no feedback. This experiment was used to assess whether the F4*_po* algorithm worked as a component of a RF algorithm, i.e. whether it provides good rankings of terms for query expansion.

The control system in this experiment only performed initial retrievals; there was no relevance feedback component of this system. The basic retrieval algorithm followed the approach given in (Ruthven, Lalmas & Van Rijsbergen, 2000b). 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, Lalmas & Van Rijsbergen, 2000b).

The experimental system performed the same search as the control system for the first query entered by the subject[10]. For the remainder of the searches on the simulated situation, each time the subject entered a query and requested a new search, a RF iteration was performed. The subject was not shown the new query terms that were added, nor were these highlighted in the full text of documents requested by the subject.

The query expansion method we used comes from (Ruthven, Lalmas & Van Rijsbergen, 2000). The systems adds to the query, for each relevant document found, the first expansion term in the expansion term ranking that appears in the document. This method provides a conservative change to the query: as few expansion terms as necessary are added to the query and the number of terms added is relative to the number of relevant documents found. This method of query expansion was shown to be generally better than adding a fixed number of expansion terms to each query (Ruthven, Lalmas & Van Rijsbergen, 2000b).

After query expansion RF systems traditionally weight query terms according to some measure of how useful they are in attracting relevant material (section 1). Our system, instead, selects 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, Lalmas & Van Rijsbergen, 2000; 2000b).

Both control and experimental systems used the same interface based on the one shown in Figure A.1. This interface did not explicitly offer a relevance feedback option; there was no *improve search* option and the subjects were only offered the *new search* option. Neither were the subjects informed that the systems were operating differently. The intention behind this decision was to be able to investigate the quality of the search mechanisms without the subject mixing search strategies (new searches and relevance feedback strategies).

The performance of a RF iteration generally takes longer than an initial search[11]. To avoid any noticeable time delay between a relevance feedback search (experimental system) and a new search (control system), which could lead the subject to avoid submitting searches on the control system, it was decided to artificially ensure that the searches took approximately the same time on both systems. For each new search (after the initial search) the control system would perform the same procedures as for an RF iteration, however the query itself was not actually modified: the RF procedures were followed but did not change the query or the way query terms were weighted. This ensures that searches on both systems took the same time to complete.

4.1 Results of Experiment One

All searches on both systems started with an initial search, subsequent search iterations on the experimental system were all feedback iterations; subsequent searches on the control system were all new searches. As we were interested only in the performance of feedback against no feedback, the information regarding the initial search was excluded and the results from Experiment One

[10] That is the first query formulation for each simulated situation.

[11] This is because the experimental system has to generate a list of candidate expansion terms, select a set of expansion terms to add to the query and select weighting schemes for each query term before running a retrieval.

only refer to the searches carried out after the initial search. This allows a direct comparison of feedback only against no feedback.

4.1.1 Overall search behaviour

The subjects overall search behaviour is summarised in Table 5. All values are average values for individual search tasks.

	Control system	**Experimental system**	**Significant**
Number of post-initial queries	**2.28**	1.56	no, *t* = 1.81
Documents viewed per simulated situation[12]	12.94	**13.67**	no, *t* = -0.18
Documents retrieved per simulated situation	**53.56**	36.33	no, *t* = 1.46
Relevant documents per simulated situation	5.41	**5.66**	no, *t* = -0.28

Table 5: Summarised search behaviour for Experiment One
Bold figures indicate highest value

Over the course of a whole simulated situation, the subjects on average viewed as many documents on both systems, and found as many relevant documents. However, on the control system, they carried out more search iterations and consequently retrieved more documents to retrieve the same number of relevant documents. Therefore it appears that the RF-selected terms plus the subjects' terms (experimental system) are more effective than the subjects' search terms alone (control system). We examine this in more detail in the next section.

4.1.2 Search effectiveness

The overall precision of the two systems, measured as the total number of unique relevant documents found divided by the total number of unique documents viewed, was roughly similar (44.52% control vs 48.48% experimental). Again these figures only relate to search iterations performed after the initial search.

Table 6 breaks these overall figures down by simulated situation. For situations 307i, 321, 322i and 347i there was an increase in precision of about 20% when using the experimental system. On situations 303i and 326i the control system gave much better performance (almost 50% increase over the experimental for situation 303i and around 24% for situation 326i).

The difference in precision between the two systems was not found to be statistically significant, (*t* = -0.31). However if we only consider the four simulated situations where the experimental system is better (307i, 321, 322i and 347i) then the experimental system is significantly better than the control system (*t* = –9.33). On the simulated situations where the control system is better (303i and 326i) the control system is not significantly better than the experimental system (*t* = 1.56).

[12] This is the number of documents viewed per simulated situation. Documents that were viewed more than once in a search are only counted once and only documents that were viewed after the initial search iteration are considered. The same applies to the count of retrieved documents (row 4).

System	**303i**	**307i**	**321**	**322i**	**326i**	**347i**
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 6: Results of documents relevant per viewed
Bold figures indicate highest value

Comparing the precision by measuring the number of relevant documents found by the number of documents *retrieved*, Table 7, it can be seen that the experimental system gives better precision for five of the six search simulated situations. Again the results overall are not significant but if we consider only the simulated situations where the experimental system is better than the control system, then the experimental system is significantly better (*t* = –4.99).

Condition	**303i**	**307i**	**321**	**322i**	**326i**	**347i**
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 7: Results of documents relevant per retrieved
Bold figures indicate highest value

Therefore the searchers are finding a higher percentage of relevant documents with the experimental system per documents retrieved and documents that the subject chooses to view. However this is not true for all simulated situations – for some situations, e.g. situation 303i, the subject performs better query modification than RF.

Finally, in Table 8 we compare the average relevance score given to the relevant documents by the subjects. For almost all simulated situations the subject gives higher scores to documents retrieved by the control system – where the subject performs the query modification. So although the experimental system, which uses RF, is better at obtaining new relevant documents it may not be better at retrieving higher quality relevant documents. The difference in relevance score was not, however, significant (*t* = 1.46).

Condition	**303i**	**307i**	**321**	**322i**	**326i**	**347i**
Control	**3.87**	**4.41**	4.78	**5.74**	**5.77**	5.25
Experimental	3.65	2.77	**5.00**	3.41	5.74	**5.41**

Table 8: Average relevance score for control and experimental system
Bold figures indicate highest value

In the next section we compare the subjects perceptions of searching on the two systems to see whether the searchers indicated a preference for one system over another.

4.1.3 Subjects' perceptions

The subjects were asked to rate certain aspects of their search, relating to their perception of each simulated situation they performed. The answers were to be given on a 5-point scale, rated from 1

(*Not at all* (useful)) to 5 (*Extremely* (useful)). For the question '*Was it easy to search on this topic?*', *'Are you satisfied with the results of your search?'*, and *'Did you have enough time to do an effective search?'* the subjects rated the experimental system higher than the control system, however the results were not significant. Table 9 summarises the differences.

	Easy to search	**Search satisfaction**	**Time to search**
Control	3.50	3.06	3.56
Experimental	**3.83**	**3.44**	**3.94**
Significant	no, t = -1.11	no, t = -1.40	no, t = -0.95

Table 9: Comparison of subject responses in Experiment One
Bold figures indicate highest value

The results from this experiment shows a preference for feedback: the searchers found the same proportion of relevant documents in searching but found these documents using less searching with the experimental system.

5 Experiment Two

The previous experiment showed that the F_4*_po* term ranking scheme could operate as part of a RF algorithm. In this experiment we compare how it performs against the standard version of F_4, i.e. whether F_4*_po* ranks terms for query expansion better or worse than F_4*_standard*. The common interface to both systems explicitly offers a RF option as well as a *new search* option. The RF option – called *improve search,* Figure A.1 – was explained to the subject as being an option which would attempt to improve the content of their query using documents they had assessed as being useful. Both systems use the same query expansion and term reweighting techniques as the control system in Experiment One, section 4. The control system uses F_4*_standard* to rank expansion terms, the experimental system uses F_4*_po*. This experiment, then, is a direct comparison between term ranking algorithms.

5.1 Results from Experiment Two

5.1.1 Overall search behaviour

In Table 10 we summarise the subjects' overall search behaviour. As can be seen the subjects carried out roughly the same number of search iterations on both systems, with a higher percentage of RF iterations on the control system. The difference between the number of new search iterations and feedback iterations on the same system was not found to be statistically significant (t = 1.83 control system, t = 1.93 experimental system), however the t values do indicate that there may be a preference, on both systems, for the subjects performing a new search over a RF search.

Overall there seems to be a preference for the control system: subjects ran more RF searches, viewed more documents per search and found more documents per search. In the next section we look at whether this also applies to the *effectiveness* of the two search systems.

	Control	Experimental	Significant
New search iterations	2.72	**2.89**	no, *t* = -0.28
RF iterations	**2.00**	1.39	no, *t* = 0.86
Documents viewed per simulated situation	**23.98**	19.67	no, *t* = 1.14
Documents retrieved per simulated situation	**101.83**	97.17	no, *t* = 0.26
Relevant documents per simulated situation	**12.89**	9.56	no, *t* = 1.29

Table 10: Summarised search behaviour for Experiment Two
Bold figures indicate highest value

5.1.2 Search effectiveness

The overall precision of the two systems indicates that the control system is more effective (precision of relevant documents to retrieved documents is 12.66% for the control system, 9.83% for the experimental system, not significant *t* = 1.20).

In Table 11, we compare the effectiveness of the two systems regarding the documents the subjects chose to view. From Table 11, we can see that, although there is no significant difference, the overall effectiveness and the effectiveness *after* RF is better on the control, F_4*_standard*, system.

	Control	Experimental	Significant
Viewed precision (relevant/viewed)	**52.15%**	49.05%	no, *t* = 0.46
Viewed precision *before* feedback	**61.55%**	60.33%	no, *t* = 0.18
Viewed precision *after* feedback	**30.03%**	18.06%	no, *t* = 0.97

Table 11: Summarised search effectiveness for Experiment Two
Bold figures indicate highest value

In Table 12, we split this down by simulated situations, looking at the average viewed precision for each of the simulated situations (relevant documents per documents viewed) after feedback. For more of the simulated situations (situations 303i, 322i, 326i and 347i) the control system gave a higher precision value. On both systems there were two situations for which no relevant documents were found after feedback.

Condition	303i	307i	321	322i	326i	347i
Control	**27.27%**	0.00%	0.00%	**41.18%**	**78.38%**	**33.33%**
Experimental	5.88%	**37.50%**	**11.76%**	0.00%	53.19%	0.00%

Table 12: Results of documents relevant per viewed after feedback
Bold figures indicate highest value

These values would appear to indicate a favour for the control, F_4_*standard*, (non partial, non ostensive) system in terms of search success. However the subjects' perceptions of the terms suggested by the system were at odds with this finding. We discuss this in the next section.

5.1.3 Subjects' perceptions

In the post-search questionnaire 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).

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_*standard* and F_4_*po* systems is shown in Figure 5. 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 5 shows the top ten terms ranked by F_4_*standard* and F_4_*po*.

F_4_***standard***	F_4_***po***
accrete	astronomer
chaisson	hubble
cullers	telescope
goldreich	universe
sandpile	astronomers
terrile	telescopes
borucki	scientists
machtley	orbit
nebula	nasa
astronomer	earth

Figure 5: Sample terms selected by F_4_*standard* and F_4_*po*

The F_4_*standard* 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 F_4_*standard*.

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, e.g. (Ruthven, Tombros & Jose, 2001). Consequently they may lose out on the potential benefits from employing RF in their searches. On the other hand, terms that appear useful to the search, even if they do not actually improve the precision of the search, may encourage subjects to interact more with the system, for example by suggesting more query terms themselves. The results of this analysis are summarised in Table 13.

In Table 13, we show the source of query terms that were added after the initial query: either added by the subject (row 3) or the system through RF (row 4). We then show how many of the terms the subject removed were those that were originally added by the subject themselves (row 5) or by the system (row 6).

	F$_4$_*standard*	**F$_4$_*po***	**Significant**
Source of added terms			
subject	2.00	**2.33**	no, *t* = -0.36
system	**3.33**	1.11	**yes**, *t* =3.78
Source of removed terms			
subject	0.72	**1.17**	no, *t* = -1.16
system	**2.28**	0.67	**yes**, *t* = 2.54

Table 13: Summary of query term addition and removal per simulated situation
Bold figures indicate highest value

Comparing the two systems, Table 13, it can be seen that the subjects on either system did not add or remove a significantly different number of their own terms (rows 3 and 5). That is, the term modification as regards their own terms were similar.

However the *system* did add a significantly different number of terms to the query (row 4). The main reason for this is that F$_4$_*po* emphasises the original query terms more than the F$_4$_*standard* algorithm. The experimental system is therefore less likely to perform query expansion and will generally modify the query less than the control system.

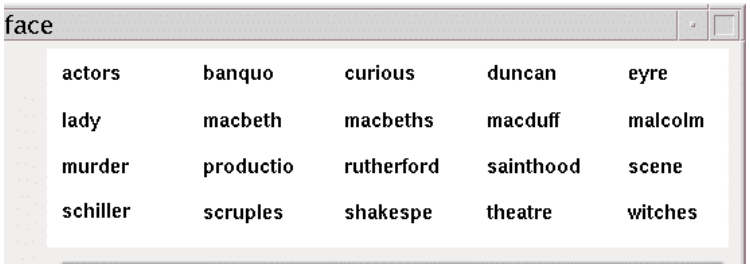
The subjects also removed a significantly higher number of the system-added terms from the query (row 6) in the control system, demonstrating that the subjects are less likely to value these terms than the ones added by the experimental system. The other difference was that the subjects were more likely to remove one of their own terms rather than a system-added one on the control system: the subjects removed 36% of their own terms and 68% of the terms suggested by the system when using the control system compared to 50% of their own terms and 60% of the system suggested terms with the experimental system.

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. We discuss possible reasons for this in section 7 but, before this, we compare the effectiveness of the two term ranking schemes when the subject is selecting new query terms – Interactive Query Expansion.

6 Experiment Three

The third experiment compared the effectiveness of the F$_4$_*standard* and F$_4$_*po* term ranking schemes in suggesting new expansion terms for selection by the subject. In this experiment the control system used the F$_4$_*standard* 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 *improve search* option, Figure A.1, was replaced by a *suggest terms* button. The expansion terms were presented in alphabetical order as in Figure 6. Clicking on an expansion term would add the term to the subject's query. Once the subject had finished query modification the subject was required to click on the *new search* button to initiate a new retrieval.



actors	banquo	curious	duncan	eyre
lady	macbeth	macbeths	macduff	malcolm
murder	productio	rutherford	sainthood	scene
schiller	scruples	shakespe	theatre	witches

Figure 6: Expansion term presentation

6.1 Results from Experiment Three

6.1.1 Overall search behaviour

In Table 14 we summarise the overall search behaviour of the searchers.

	Control	Experimental	Significant
Search iterations	**4.22**	4.17	no, *t* = 1.07
Documents viewed per simulated situation	23.17	**25.78**	no, *t* = -0.15
Documents retrieved per simulated situation	59.22	**66.72**	no, *t* = 0.51
Relevant documents per simulated situation	9.22	**12.72**	no, *t* = -1.35

Table 14: Summary of overall search behaviour for Experiment Three
Bold figures indicate highest value

From Table 14 it can be seen that the subjects performed roughly the same number of searches per situation and they tended to search in a similar fashion; retrieving and viewing roughly the same number of documents per situation.

Our main interest is this experiment, however, is how the subjects utilised the terms suggested by the control and experimental systems. In Table 15 we present details on the source of terms that were added or removed by the subjects. In particular we count the number of added/removed terms that were generated by the user and the number of added/removed terms that were suggested by the system.

From Table 15, it can be seen that with the control system the subject was more likely to add their *own* terms to their query than ones suggested by the system, (on average per situation subjects added 8.83[13] of their own terms compared against 1.61 of the expansion terms suggested by the system). On the experimental system, however, this was reversed: the subject was more likely to add terms suggested by the system (8.17 terms per search, compared against 6.67 of their own).

[13] This does not include the original query terms.

The difference between the number of their own terms the subject 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 suggestion algorithm.

Control	**303i**	**307i**	**321**	**322i**	**326i**	**347i**	**Averages**
Own terms added by subject	26	8	**26**	**20**	**64**	15	**8.83**
System suggested term added by subject	4	2	9	4	4	6	1.61
Own terms removed by subject	16	**6**	**29**	**18**	**63**	0	**7.33**
System suggested term removed by subject	1	**2**	**9**	**1**	2	0	**0.83**
Experimental	**303i**	**307i**	**321**	**322i**	**326i**	**347i**	**Averages**
Own terms added by subject	**31**	**14**	**26**	16	11	**22**	6.67
System suggested term added by subject	**36**	**12**	2	**29**	**33**	**35**	**8.17**
Own terms removed by subject	**20**	4	23	2	10	**10**	3.83
System suggested term removed by subject	**2**	0	2	0	**6**	**0**	0.56

Table 15: Statistics on query terms in Experiment Two per simulated situation
Bold figures indicate highest value

The subjects also tended to remove fewer expansion terms, either those suggested by the system or themselves, with the control system. Neither difference here was significant (difference in subject-suggested terms removed *t* = 1.14, difference in system-suggested terms *t* = 0.56).

6.1.2 Search effectiveness

The previous section showed that subjects tended to use more terms suggested by the F_4_*po* term ranking scheme used by the experimental system. In this section we investigate whether the increase in term use lead to an increase in retrieval effectiveness: did using more expansion terms lead to the retrieval of more relevant documents?

In Table 16 we present the number of unique relevant documents found on average per simulated situation and the average relevance score given by the subjects to the documents they assessed as relevant. From Table 16, it can be seen that on all situations, with the exception of situation 321, 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 suggest 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. More experimentation here is needed to validate or clarify this claim.

		303i	**307i**	**321**	**322i**	**326i**	**347i**
Control	Relevant documents found	10.00	**8.00**	**12.33**	7.33	9.67	8.00
Experimental	Relevant documents found	**11.00**	**8.00**	7.00	**9.33**	**21.67**	**9.33**
Control	Average relevance score	3.78	5.37	5.14	5.05	4.49	4.31
Experimental	Average relevance score	**6.91**	**6.82**	**6.01**	**7.33**	**7.08**	**5.48**

Table 16: Comparison of relevant documents found and average relevance score
Bold figures indicate highest value

6.1.3 Subject's perceptions

The subjects were asked to rate certain aspects of their search relating to their perception of each simulated situation they performed. Table 17 summarises the subject's perceptions of the search as they relate to the expansion term suggestions. In particular we concentrate on the results to the questions '*Was it easy to search on this topic?'*, *'Are you satisfied with the results of your search?'*, *'Did you have enough time to do an effective search?'* and *'How useful do you think the query words, suggested by the system, were to your search?'*. All responses were on a scale of 1-5 with a score of '1' representing the category *'Not at all*' and a score of '5' representing the category *'Extremely*'.

	Easy to search	Search satisfaction	Time to search	Utility of terms
Control	2.72	2.61	3.33	1.53
Experimental	**3.72**	**3.83**	**3.89**	**3.53**
Significant	no, t = -1.72	**yes**, t = -2.99	no, t = -1.41	**yes**, t = -3.73

Table 17: Comparison of subject responses in Experiment Three
Bold figures indicate highest value

For all questions the subjects rated the experimental system higher: they found it easier to perform searches upon, had higher search satisfaction and were generally happier with the time they were given to search. More importantly, the subjects rated the terms suggested by the experimental system as better than those suggested by the control system. This, and the subjects' satisfaction with their search, was significantly different in favour of the experimental system.

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 subject and which lead to the retrieval of better relevant documents.

7 Summary of incorporating user behaviour into term ranking

The aim of this part of the paper was to investigate how user search behaviour could be incorporated into the term ranking process. In section 3 we used the standard F_4 term ranking function to incorporate partial relevance assessments and ostensive evidence. That is, we were using information on *when* the documents were marked relevant and *how* relevant the users regarded the documents. Therefore we were using information on the relevance assessments themselves, rather than just the content of the relevant documents, to influence the term ranking process.

Three experiments were carried out to investigate the effectiveness of the new term ranking function, F_4*_po*. In Experiment One we showed that the F_4*_po* can act successfully as part of a RF algorithm: searchers found the same proportion of relevant documents in fewer search iterations than without RF.

In Experiment Two we compared the performance of F_4*_po* against F_4*_standard* for automatic query expansion. We showed that, although, the use of the F_4*_po* term ranking function did not increase the retrieval effectiveness, the subjects' perceptions were that the new version, F_4*_po,* provided more useful terms. This experiment is interesting in the lack of correlation between the subjects' perceptions (their view of the expansion terms), their interaction with RF (the fact that they appeared to use the F_4*_po* terms more and remove them less often) and with how useful the documents retrieved by these terms were. That is, although the subjects liked the F_4*_po* terms they did not necessarily lead to the retrieval of more relevant documents.

There are two possible reasons for this. One reason is that we did not weight the expansion terms relative to the original query terms. This has previously been shown in test collection evaluations, e.g. (Haines & Croft, 1993; Salton & Buckley, 1990), to be a good technique to avoid the expansion terms dominating the retrieval of new documents. We did not employ this technique in these experiments as we lacked a good method of estimating weights for the original and expansion terms relative to each other. During the post-search interview several subjects mentioned that RF, using the F_4*_po* system, seemed to retrieve too few documents containing the original query terms.

The second reason that the F_4*_po* system may perform less well is that, on average, the query expansion technique, section 4, emphasises a minimal change to the content of the user's query. Although this technique is generally more effective than adding a fixed number of terms (Ruthven, Lalmas & Van Rijsbergen, 2000) it tends to add fewer terms when F_4*_po* is used to rank the terms, as described in section 5. Therefore the F_4*_po* system in this experiment generally adds *fewer* terms to the query, Table 13, possibly reducing the effectiveness of the F_4*_po* system. More experimentation is needed to investigate this.

In Experiment Three we showed that, in an interactive query expansion situation, the subjects[14] not only preferred the F_4*_po* term suggestions but added more of them to their query and assessed the retrieved documents as having higher relevance. In both Experiments Two and Three, the subjects seemed to trust the F_4*_po* terms more than the ones suggested by F_4. The incorporation of user search behaviour, then, can have positive effect on RF algorithms. In the remainder of this paper we examine the role of query expansion.

8 User behaviour and query expansion

The previous sections concentrated on ordering terms prior to query expansion; in the following sections we shall investigate the techniques for query expansion itself. In this section we shall describe how we use search behaviour to aid the choice of expansion terms. This is a summarised account of our previous work on test collections (Ruthven, 2001).

The central argument in (Ruthven, 2001) was that it is possible to use searcher behaviour as an indication of what *type* of query expansion should be used for an individual search. That is, some types of query expansion are more successful for some types of retrieval situation. For example, low precision searches are often better handled by an expansion technique that gives a larger change to the content of the query, and high precision searches are better handled by an expansion technique that is more selective about which terms are added to the query.

We also showed that it was possible to *choose* which query expansion technique was likely to be most effective for individual queries based on information such as the precision of a search, where in the ranking the relevant documents appeared and the similarity of the relevant documents within a search. This type of evidence has previously been shown by e.g. Spink et al (Spink, Greisdorf & Bateman, 1998), and Vakkari, (Vakkari, 2000), to be an important indicators of the user's search process.

[14] A separate set of subjects from Experiment Two.

What is important about this approach to RF is that we use interactive features of the search to select RF techniques; the system adapts RF to information on how a user is searching. Our approach to choosing a query expansion technique is based on analysing interactive features of the search. We concentrate on the three following pieces of evidence:

i. precision of the search, i.e. how many relevant documents have been found in the search before the user initiates feedback?
ii. position of documents within the document ranking, i.e. where in the document ranking are the relevant documents found?
iii. similarity of relevant documents, i.e. how similar are the relevant documents to each other, and how similar are the documents to the retrieved (non-relevant) documents?

In (Ruthven, 20001) we showed that the values of these three attributes can be used to select a query expansion technique; these can either be techniques that force a minimal change in the content of the user's query, e.g. coverage (one that emphasises the terms that are similar in the relevant documents), or Josephson, (one that emphasises the discriminatory terms) or they can be query expansion techniques that force a larger change in the content of the query (maximal expansion). In our experiments a maximal expansion corresponded to the addition of the top six expansion terms.

The evidence can be used to calculate a set of rules, Figure 7, for selecting which type of query expansion is required at any iteration of RF. Each rule provides support for one query expansion technique and the expansion technique with the highest support is selected to perform RF (Ruthven, 2001). In (Ruthven, 2001) we showed that this technique can be effective within a test collection evaluation. In this paper we investigate whether it works for interactive searches.

```
if (term ranking method = F4_po)
    if (precision is high) use josephson
        else if (precision is low) use maximal
    if (position is high) use coverage
        else if (position is low) use maximal
    if (similarity is high) and (number of relevant documents is high) use coverage
        else if (similarity is high) and (number of relevant documents is low) use josephson
        else if (similarity is low) use maximal
```

Figure 7: Rules for selecting query modification technique for the F_4_*po* term ranking scheme where **bold** entries indicate features of the retrieval, *italic* entries indicate values of the features, and underlined entries indicate the query modification techniques suggested by the value of the feature

9 Experiment Four

In Experiment Four we compare this process of selecting query expansion techniques, outlined in section 8, against using a single query expansion technique for all iterations of RF. In Experiment Four, the control system adds the top six expansion terms to the query for each iteration of RF. Each iteration of RF, therefore, uses the same algorithm for query modification. The experimental system *selects* which RF technique to use based on the behavioural evidence given by the searcher as outlined in the previous section. Both systems use the same interface.

9.1 Overall search behaviour

In Table 18 we summarise the main findings from the subjects interaction with the two systems. On the experimental system the subjects carried out more new searches, more RF and viewed more documents than on the control system. On the control system the subjects found more relevant documents. However there were no significant differences in search behaviour.

	Control	Experimental	Significant
New search iterations	2.34	**2.89**	no, *t* = -1.98
RF iterations	1.06	**1.17**	no , *t* = -0.79
Documents viewed per search task	16.95	**19.22**	no, *t* = -1.46
Documents retrieved per search task	57.89	**61.34**	no, *t* = -0.93
Relevant documents per search task	**9.56**	8.39	no, *t* = 1.29

Table 18: Comparison of searches on control and experimental system
Bold figures indicate highest value

9.2 Search effectiveness

The overall precision of the control system was higher than the experimental system whether it is measured as the relevant documents found compared against the number of documents the subject viewed (54.80% control, 46.98% experimental) or against the number of documents retrieved (17.90% control, 14.82% experimental). Neither of these differences were significant (*t* = 0.85 viewed documents, *t* = 1.09 retrieved documents).

In the remainder of this section we compare the results only for RF iterations: the results of searches that were initiated by the subject selecting the RF option. This will give a clearer picture of the relative performance of the two RF techniques used in this experiment.

After feedback the subjects had relatively similar precision values, as measured by the number of documents found after feedback divided by the number of documents viewed after feedback (50.78% control, 52.08% experimental). The results are not significant (*t* = -0.07) and for two situations the control system gives better precision whereas the experimental system gives better precision for the other four situations, Table 19.

Condition	303i	307i	321	322i	326i	347i
Control	63.19%	**100.00%**	18.26%	**59.88%**	24.81%	38.57%
Experimental	**70.01%**	42.18%	**80.07%**	19.94%	**36.81%**	**63.49%**

Table 19: Precision of documents relevant per viewed after feedback
Bold figures indicate highest value

In Table 20, we show the average relevance score for documents after a new search, after RF, and the ratio of the scores after and before feedback. This latter measure gives an indication of whether the documents found after RF are given higher relevance scores than after a new search. A value of greater than one indicates higher relevance scores after RF and a value of less than one indicates lower relevance scores after feedback.

From Table 20 it can be seen that, on average, the relevance scores for the experimental system are higher than the control system for new search and after RF. However the ratio measures are virtually identical. This shows that, although, we achieve higher relevance scores with the experimental system, the experimental system does not retrieve better relevant documents after RF than it was retrieving after a new search.

	Average		Average
Control system before RF	5.44	**Experimental system before RF**	**5.63**
Control system after RF	3.73	**Experimental system after RF**	**4.47**
Ratio before/after	0.76	**Ratio before/after**	**0.78**

Table 20: Ratio of relevant scores before and after feedback
Bold figures indicate highest value

9.3 Subject's perceptions

In this section we compare the subjects' perceptions of the two systems. In particular we concentrate on the subjects' responses to three aspects: their satisfaction with the search, their assessment of whether they had sufficient time to search and their assessment of how useful RF was to their search.

Question	Control system	Experimental system	Significant
Search satisfaction	**3.72**	3.33	no, *t* = 0.97
Time for search	3.50	**3.67**	no, *t* = -0.59
Utility of RF	1.72	**3.01**	**yes**, *t* = -3.50

Table 21: Average subject responses in Experiment Four
Bold figures indicate highest value

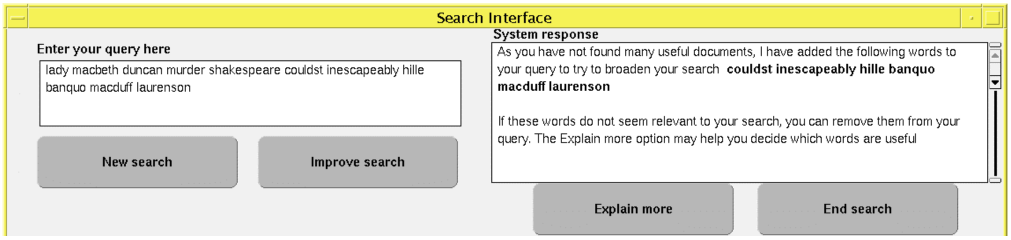
In Table 21 we present the average response to these questions and whether the difference is significant. As can be seen the results are not conclusive in favour of one or other systems: the subjects had greater satisfaction with the control system but felt they had less time with this system and rated the RF component lower than the experimental system. This set of results is important because they do not show a major difference: the systems different methods of choosing expansion terms but there was no noticeable performance difference between the two systems. However again there was a preference for the experimental system.

10 Experiment Five

The fifth experiment concentrates on the role of RF at the interface. One of the reasons the subjects reported being unwilling to use RF was the poor relation between the effects of RF and their search. That is they were not sure how RF worked, what it was going to do to their search or how to undo the effects of RF. In this experiment we developed a new interface to test whether giving the user more information on the effect of RF would increase the use of RF.

The control system is identical to the experimental system from Experiment Four.

The experimental system uses the same RF technique but uses a different interface[15]. The interface for the experimental system contains the same components as the control system except that, each time the subject issues a RF request the system presents an explanation of the effect of RF on their search, Figure 8. A more detailed account of the structure and intention behind the creation of explanations can be found in (Ruthven, 2002).



lady macbeth duncan murder shakespeare couldst inescapeably hille banquo macduff laurenson

As you have not found many useful documents, I have added the following words to your query to try to broaden your search **couldst inescapeably hille banquo macduff laurenson**

If these words do not seem relevant to your search, you can remove them from your query. The Explain more option may help you decide which words are useful

Figure 8: Explanation of RF

The explanations are of five types; three of which correspond to the types of query expansion outlined in section 8 (**i** – **iii**), the remaining two deal with cases where no query expansion occurs (**iv** and **v**):

i. *maximal explanation*. In this case the user has marked few documents relevant and the system attempts to broaden the user's search by adding more search terms. The system lists the terms that it has added and displays a message like this *'As you have not found many useful documents, I have added the following words to try to broaden your search* ***couldst inescapeably hillle banquo macduff laurenson****'*. In this example, ***couldst inescapeably hillle banquo macduff laurenson*** are the top six expansion terms.

ii. *coverage explanation*. In this case the system will present the user with an explanation like this '*I have added the words* ***macduff banquo*** *to your query as they appear in most of the documents you have marked useful*'. This type of explanation emphasises the search terms that make the user's documents similar to each other.

iii. *Josephson explanation*. In this case the system will present the user with an explanation like this '*I have added the word* ***macduff banquo*** *to your query as they appear to be important to your search'*. This type of explanation emphasises search terms that are good discriminators of relevance.

iv. *no expansion explanation.* As described in section 4 the F_4*_po* term ranking function tends to place the original query terms high up the list of expansion terms. If the original query terms are judged to be the best terms then the system will not add any search terms to the user's query but instead will concentrate on improving the weighting of the search terms – selecting good term and document weighting schemes. The explanation presented at the interface therefore concentrates on how the query terms are weighted rather than which terms are used. A sample explanation of this type is '*Based on the documents you have marked useful, I will treat* ***macbeth*** *as the most important word in your search and try to retrieve more documents containing this word'*.

[15] This is the only experiment in which the interfaces for the control and experimental system differ.

v. *don't know explanation*. If the system cannot choose one good explanation – the evidence available is split between different query expansion types for example – then the system will tell the user it cannot decide what kind of documents the user requires. It will show the user a message suggesting the user provides more evidence. For example, ''*I am not sure what kind of documents you want – perhaps you could mark some more documents as useful or add some more words to your query. Here are some examples that may be useful* ***banquo theatre macduff king scene arts****'*. As in the expansion explanation, **i.**, the terms ***banquo theatre macduff king scene arts*** are the top-ranked expansion terms.

In all explanations the system also offered some suggestion on how the user could change, correct or improve the system's decision, e.g. removing poor terms, adding more similar terms, marking more documents relevant, (Ruthven, 2002).

The user can also explicitly request more information on the RF process. This is by means of an *Explain more* option, Figure 8. This option will expand the information contained within the explanation with information on how terms are used to select the new set of retrieved documents. The *Explain more* option can give three types of information regarding each search term.

i. It can tell the user which terms are being treated as important to the main topic of the document. The system presents a message like '*I am looking for documents where* ***macduff*** *seems important to the main topic of the document'*.

ii. It can tell the user which terms should appear often in retrieved documents. In this case the system will present a message like '*I am looking for documents that contain lots about* ***macduff'***.

iii. If simply the presence of the term is important, the system will simply tell the user that these terms are important, e.g. '*I am looking for any documents that contain the word* ***macduff'***.

In this experiment we look at the effectiveness of these summaries in helping subjects to understand what effect the RF algorithm is having on the search. Unlike the other experiments, the control and experimental systems differed at the interface rather than the underlying system. Therefore the main focus in the following sections is to highlight the main differences in the two systems regarding how the overall system was used rather than the effectiveness of the RF engine itself.

10.1 Overall search behaviour

In Table 22 we summarise the overall search behaviour regarding the control (no explanation) system versus the experimental (explanation) system. In this experiment unlike the other experiments there were noticeable differences in the search behaviour; the subjects performed significantly more RF searches, viewed more documents and retrieved more documents using the experimental system.

	Control	**Experimental**	**Significant**
New search iterations	**2.11**	**2.11**	no, *t* = 0.00
RF iterations	1.50	**1.95**	**yes** , *t* = -3.16
Documents viewed per search task	19.56	**26.06**	**yes**, *t* = -2.58
Documents retrieved per search task	46.44	**59.72**	**yes**, *t* = -3.33
Relevant documents per search task	**9.78**	9.61	no, *t* = 0.14

Table 22: Comparison of searches on control and experimental system
Bold figures indicate highest value

In Table 23 we compare how often subjects performed a new search with how often they performed RF. The subjects on both systems tended to run more new searches than RF searches on both systems. This is in line with previous investigations of subject use of RF, e.g. (Beaulieu, 1997), although if we exclude the initial search iteration, as each subject had to issue at least one query per simulated situation, then subjects can be seen to be performing slightly more RF than new searches.

From Table 23 the subjects, on average, performed the same number of new searches on both systems. However they tended to perform more RF searches on the experimental system.

The number of new search iterations performed on the two systems was not statistically significant. The difference in number of RF iterations, however, was found to be statistically significant (*t* = 3.16). Therefore the subjects were running more RF iterations on the experimental system. The *percentage* of all search iterations that were RF iterations (Table 23 rows 5 and 10) was not significant (*t* = -0.92). These two results indicate that, although there was no preference for using RF over new searches on either system, there *was* a preference for using RF on the experimental system.

	303i	**307i**	**321**	**322i**	**326i**	**347i**	**Average**
Control system							
New search iterations	3.00	2.00	1.00	2.67	2.67	1.33	**2.11**
RF iterations	1.67	1.33	1.67	1.33	1.33	1.67	1.50
%age RF iterations/new search iterations	36%	40%	**63%**	33%	**33%**	56%	43%
Experimental system							
New search iterations	2.00	2.33	2.00	1.33	3.67	1.33	**2.11**
RF iterations	2.00	2.00	1.67	2.33	1.67	2.00	**1.95**
%age RF iterations/new search iterations	**50%**	**46%**	46%	**64%**	31%	**60%**	**49%**

Table 23: Comparison of new searches against RF searches
Bold figures indicate highest value

10.2 Search effectiveness

In Tables 24 and 25 we present the precision of documents assessed relevant to the number of documents viewed by the subject (Table 24) and the precision of documents assessed relevant to the number of documents retrieved (Table 25).

Condition	303i	307i	321	322i	326i	347i
Control	46.55%	**32.64%**	**42.65%**	**53.35%**	46.98%	**37.03%**
Experimental	**62.25%**	16.48%	37.50%	20.55%	**52.49%**	31.55%

Table 24: Precision of documents assesed relevant per documents viewed
Bold figures indicate highest value

Condition	303i	307i	321	322i	326i	347i
Control	15.76%	**12.69%**	**26.54%**	**21.78%**	20.21%	**16.16%**
Experimental	**19.30%**	6.52%	17.76%	11.22%	**20.90%**	12.62%

Table 25: Precision of documents assesed relevant per documents retrieved
Bold figures indicate highest value

In both Tables 24 and 25 the experimental system gave better performance for situations 303i and 326i, whereas the control system gave better performance on the other four situations. In neither case was the difference significant (*t* = 1.78 retrieved documents, *t* = 0.93 viewed documents). If we only consider the situations where the control system performed best, the control system was significantly better than the experimental system in precision of relevant documents per retrieved documents (*t* = 4.74).

There is a preference for the control system in terms of these performance measures. This is because in both cases, although the subjects found more relevant documents with the experimental system, Table 22, they had to view more documents and retrieve more documents with the experimental system to obtain the same number of relevant documents.

10.3 Subject's perceptions

In Table 26 we summarise the subjects' overall perceptions of their search. There were no significant differences between the two systems, in particular there were similar ratings for the RF function of the two systems. This means that the subjects do not perceive RF as more useful on either system. However, as shown in section 10.1, the subjects do *use* RF more on the experimental system.

Question	Control system	Experimental system	Significant
Search satisfaction	**3.50**	3.33	no, *t* = 0.47
Time for search	**3.45**	3.00	no, *t* = 0.90
Utility of RF	**3.83**	**3.83**	no, *t* = 0.00

Table 26: Average subject responses in Experiment Four
Bold figures indicate highest value

An important aspect of this experiment is whether the use of explanations helped the subjects understand RF and to what degree they stimulated the subjects' interest in RF. In particular we examine how useful the subjects rated the three features: RF, the explanation itself and the *Explain more* option.

In Table 27 we compare the average subject score for the three options. Each subject was asked how useful the options were to their search. As in previous questions the subject was asked to indicate the utility of the option using a 5 point scale with the value of '5' reflecting the highest utility. The values shown in Table 27 show the averaged results for the searches in which a subject employed RF[16].

The general tendency is for the Explanation to be rated higher than the RF option which, in turn, is rated higher than the *Explain more* option. The only significant difference between the results was between the worth of the Explanation and *Explain more* option (RF vs Explanation t = -1.69, RF vs *Explain more* t = 0.79, Explanation vs *Explain more* t = 2.94).

Situation	**RF**	**Explanation**	**Explain more**
303i	3.00	**3.33**	3.00
307i	**2.50**	**2.50**	2.00
321	2.33	**3.00**	2.00
322i	1.67	**3.33**	3.00
326i	**3.33**	3.00	2.50
347i	1.50	**2.00**	0.00[17]

Table 27: Comparison of subject responses in Experiment Five
Bold figures indicate highest value

The post-search interview was used to elicit the subject's perceptions on the relative worth of these options. The main reason given for the higher rating for the Explanation was that even if RF did not work, i.e. added unhelpful terms to the query, or if the wrong type of documents were retrieved the Explanation still gave useful information. This is because it still gives information on *why* the system modified the query. Therefore the success of the Explanation is not dependent on the success of RF.

The *Explain more* option was generally rated lower than the RF option. There are two reasons for this. Firstly, subjects had to explicitly request more information. This meant that subjects may not have requested information that could have been useful if they had viewed it. Secondly, the information provided by the *Explain more* option was only useful relative to what was provided by the Explanation and RF: if the Explanation was not useful or RF led to a poor change in the subject's query then the *Explain more* option was not useful. This is because *Explain more* in this case gave more information about an aspect of the system that was not of interest. In addition, if the Explanation gave enough information to the subject about the effect of RF then the *Explain more* option was not necessary.

The situation where *Explain more* was most useful was where the subject was unsure why a query had retrieved a particular set of documents. In this case the subject could investigate the *Explain more* information to check what weighting schemes the system was using to retrieve documents. Although the subject could not change the retrieval scheme themselves they could remove terms from the query that were being prioritised by the system. A natural extension to the interface would be to allow the subject to alter the way terms were being used to retrieve documents. Overall the subjects found the *Explain more* option interesting but not always of use.

In general the subjects liked the use of explanations but most said that they would like more types of explanations and explanations that were more specific to their search. The first comment is

[16] We also checked for possible relationships between the subjects' perceptions of the options and their perceptions of how easy the simulated situation was and the success of RF. We found no detectable correlation.
[17] No subject used the *Explain more* option for this situation.

valid and a wider range of explanations could be developed for such an interface. The second comment specifically relates to the selection of query terms. Most subjects who made this comment would have preferred a more semantic explanation of why a particular query term(s) was added to their query, e.g. an explanation of the form '*I am adding the word* **space** *to your query as you are searching for documents on the Hubble telescope and space is a word that is strongly related to this topic'*. This type of explanation is very difficult to create using the statistical techniques that underlie most statistical RF systems. Most subjects liked the presentation of explanations on the basis that *some* form of system explanation was useful and encouraging. As mentioned before this was because explanations can be helpful even when RF is not performing correctly.

11 Discussion

In this section we discuss the major findings and implications of these experiments. Our overall intention was to investigate whether evidence given by a searcher whilst interacting with an information retrieval system could be used to influence relevance feedback algorithms. There are two main reasons why we are interested in this area. Firstly, users often give very little information to the system in the form of relevance assessments, i.e. they often give the system very few examples of what kind of documents they want retrieved, e.g. (Ruthven, Tombros & Jose, 2001). However, searchers, through the process of making relevance assessments – selecting which documents to view, viewing the document, assessing how relevant is the document - can are implicitly giving information to the system on the process of searching. An argument for being interested in search behaviour, then, is to try to give the system a wider range of evidence upon which to base query modification decisions.

Secondly, users can change what kind of information they require from a system. Many researchers, e.g. Vakkari (Vakkari, 2000; 2000b), Kuhlthau (Kuhlthau, 1991; 1993) and Ellis (Ellis, 1989) have shown that users interact differently at different stages in a search. Further indications of how a user is interacting with a system can be a useful indication of what the user wants the system to do (Kuhlthau, 1991). Trying to map a user's search behaviour to appropriate responses by the system can help make IR systems more responsive to changes in a user's interaction styles while searching.

We examined using aspects of user search behaviour for two functions; ranking possible new expansion terms for query expansion and deciding how to choose which expansion terms to add to the query. These investigations concentrated on different aspects of search behaviour.

In the set of experiments, Experiments One-Three, we looked at the role of behavioural evidence in the term ranking process for query expansion. In this part of the paper we examined information on the relevance of a document from two aspects; explicit evidence coming from the user's assessment of *how* useful a document was to their search, and implicit evidence coming from when the document was assessed relevance. What we were trying to demonstrate here was that we could extend the evidence for term ranking from the *content* of the relevant document (the terms it contains) to information on the relevance assessment itself (how the assessment of relevance contributes to the overall search). Evidence such as time, and degree of relevance, allow the system to prioritise the relevance evidence coming from the user. Other possible aspects of relevance assessment that could be handled in this way include factors such as the order in which users assess relevance within a search iteration. This has previously been shown to be an important differentiator of what documents a user regards as being important to their search (Florance and Marchionini, 1995). In our experiments, we included partial and ostensive evidence as an extension to the standard F_4 term ranking function could be extended to incorporate partial relevance assessments and ostensive evidence. Extensions to other algorithms, such as Porter's algorithm (Porter & Galpin, 1988), or Robertson's *wpq* algorithm (Robertson, 1990) are also possible in a similar manner.

In Experiment One we showed that the terms suggested by our new method of ranking expansion terms, F_4*_po*, allowed the users to retrieve more relevant documents in fewer search iterations. In Experiment Two, we showed that the F_4*_po* algorithm did not perform better, in the sense of retrieval effectiveness, than the baseline F_4 algorithm. In particular it did not lead to the retrieval of more relevant documents. However this may be a factor of the particular query expansion technique we used. The main difference between the two term ranking algorithms was the perception of the subjects regarding how useful the terms were. In Experiment Two, and Experiment Three which compared the two algorithms for interactive query expansion, subjects not only rated the terms suggested by F_4*_po* as being more useful but used these terms more heavily. For example in Experiment Three the subjects used more expansion terms in their query modifications. The subjects also displayed a greater degree of trust in these terms demonstrated by their willingness to leave the terms in the query for future search iterations. The subject's perception of terms is important; if subjects do not regard RF as useful then subjects may not use RF and miss out on a useful technique for searching.

In the first set of experiments we concentrated primarily on individual documents; in the second set of experiments we looked at differences in the *set* of documents assessed relevant during a search iteration. In these experiments, Experiments Four and Five, we examined the role of behavioural evidence in the query expansion stage; selecting which terms to add to the user's query. Specifically we look at how aspects of the user search behaviour such as the number of documents assessed relevant, and the similarity of the relevant documents can be used to select which query expansion technique to apply to the query[18]. Put simply, what we are trying to do is gather information on *how* the user is interacting with the results of a search and choose the most appropriate method of modifying the user's query for that particular search situation.

In Experiment Four we compared this method of selecting query expansion techniques against a standard method of selecting terms, expansion by a fixed number of terms. The results from this experiment were largely inconclusive. That is, although the subjects *perceived* RF as being more effective, it did not appear to lead to the retrieval of more relevant documents. A more detailed investigation of this experiment is needed to uncover the cause of this finding. One possible method of improving the performance of the system is to have more methods of detecting what aspects of searching is important, and a more detailed examination of how these should be used for query modification. The work on information-seeking by, e.g. Kuhlthau may form the basis of such an investigation (Kuhlthau, 1991).

In Experiment Four the experimental system examined how the user was interacting with the system and choose a query expansion technique, based on the user's style of interaction. However, the users were *not* aware that the system was performing these functions. That is, users only saw the results of query modification, they were given no insight into why or how the query had been changed.

In Experiment Five, we attempted to give the user some insight into how the system used the behavioural information it detected within the user's search. Specifically the system we

[18] The mechanics of how we decide which techniques to apply are described elsewhere, (Ruthven, 2001).

developed attempted to *explain* to the users why RF had changed their search. It gave the user reasons for changing their search, based on the user's interaction. The presentation of explanations in this way led to a higher *use* of RF. That is, the explanations could encourage the subjects to use RF by making RF a more approachable system function. The experimental system used in Experiment Five was the system that appeared most responsive to users. The system presented RF in such a way as to connect what the system was looking for (the query terms) with how the users were searching (the aspects of search behaviour) and connect both these aspects to the system's actions (query modification). The potential benefit of this approach, which needs further investigation, is that we can help the users interact better with RF systems. If users can see the consequence of their actions then they may gain more understanding of what are good actions within the context of a search.

In summary, what we are trying to do is connect how users search with how RF decisions are made to gain both more flexible systems and more personalised responses from these systems. These investigations are preliminary, the experiments are limited by their size for example, however the results do indicate that user search behaviour *can* give useful results and is an area that should be more fully exploited in IR system design.

12 Acknowledgements

We are very grateful for the helpful comments contributed by Pia Borlund and Mark Dunlop on the design of our experiments, and for the many comments made by the anonymous referees. We would also like to thank the subjects that participated in our experiments and the Library and Information Commission who funded this work in the *Retrieval through explanation* project. More information about this project can be found at http://ir.dcs.gla.ac.uk/projects/explanation/

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. pp 71-90. 2000.

Borlund, P. *Evaluation of interactive information retrieval systems*. PhD Thesis. Abo Akademi University. 2000b.

Borlund, P. & Ingwersen, P. *The application of work tasks in connection with the evaluation of interactive information retrieval systems: empirical results*. Mira '99. S. Draper, M. Dunlop, I. Ruthven and C. J. van Rijsbergen (eds). Electronic Workshops in Computing. British Computer Society. 1999.

Campbell, I. *Interactive evaluation of the Ostensive Model, using a new test-collection of images with multiple relevance assessments*. Journal of Information Retrieval. **2**. 1. pp 89-114. 2000.

Campbell, I. & van Rijsbergen, C. J. *Ostensive model of information needs.* Proceedings of the Second International Conference on Conceptions of Library and Information Science: Integration in Perspective (CoLIS 2). Copenhagen. 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.

Ellis, D. *A behavioural approach to information system design*. Journal of Documentation. **45**. 3. pp 171-212. 1989.

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

Haines, D. & Croft, W. B. *Relevance feedback and inference networks.* Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp 2-11. Pittsburgh. 1993.

Harman, D. *Relevance feedback and other query modification techniques.* Information Retrieval: Data Structures & Algorithms. Englewood Cliffs: Prentice Hall. (W.B. Frakes and R. Baeza-Yates ed). Chapter 11. pp 241-263. 1992.

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. 1992b.

Kuhlthau, C. C. *Inside the search process: information seeking from the user's perspective*. Journal of the American Society of Information Science. **42**. 5. pp 361 - 371. 1991.

Kuhlthau, C. C. *Principle for uncertainty for information seeking*. Journal of Documentation. **49**. 4. pp 339-355. 1993.

Over, P. *TREC-6 interactive track report*. Proceedings of the Sixth Text Retrieval Conference. Nist Special Publication 500-240. Gaitherburg. pp 73-82. 1998.

Porter, M. & Galpin, V. *Relevance feedback in a public access catalogue for a research library: Muscat at the Scott Polar Research Institute*. Program. **22**. 1. pp 1 - 20. 1988.

Robertson, S. E. *On term selection for query expansion.* Journal of Documentation. **46**. 4. pp 359-364. 1990.

Ruthven, I. *Abduction, explanation and relevance feedback*. PhD thesis. University of Glasgow. 2001.

Ruthven, I. *On the use of explanations as a mediating device for relevance feedback*. Proceedings of the Sixth European Conference on Digital Libraries. ECDL 2002. Lecture Notes in Computer Science. Rome. 2002.

Ruthven, I., Lalmas, M. & van Rijsbergen, C. J. *Empirical investigations on query modification using abductive explanations*. Proceedings of the Twenty-Fourth ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans. pp 181-189. 2001.

Ruthven, I., Lalmas, M. & van Rijsbergen, C. J. *Combining and selecting characteristics of information use*. Journal of the American Society for Information Science and Technology. **53**. 5. pp 378-396. 22001b.

Ruthven, I., Tombros A, & Jose, J. *A study on the use of summaries and summary-based query expansion for a question-answering task* Twenty-Third BCS European Annual Colloquium on Information Retrieval Research (ECIR 2001). Darmstadt. 2001

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

Robertson, S. E. & Sparck Jones, K. *Relevance weighting of search terms.* Journal of the American Society for Information Science. **27**. 3. pp 129-146. 1976.

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

Vakkari, P. *Cognition and changes of search terms and tactics during task performance.* Proceedings of RIAO Conference on Content-Based Multimedia Information Access. Paris. pp 894-907. 2001.

Vakkari, P. *Relevance and contributing information types of searched documents in task performance*. Proceedings of the Twenty-Third ACM SIGIR Conference on Research and Development in Information Retrieval. pp 2-9. Athens. 2000b.

Voorhees, E. H. & Harman, D. *Overview of the sixth text retrieval conference (TREC-6).* Information Processing and Management. **36**. 1. pp 3-35. 2000.

Appendix

Simulated situation 303i
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.

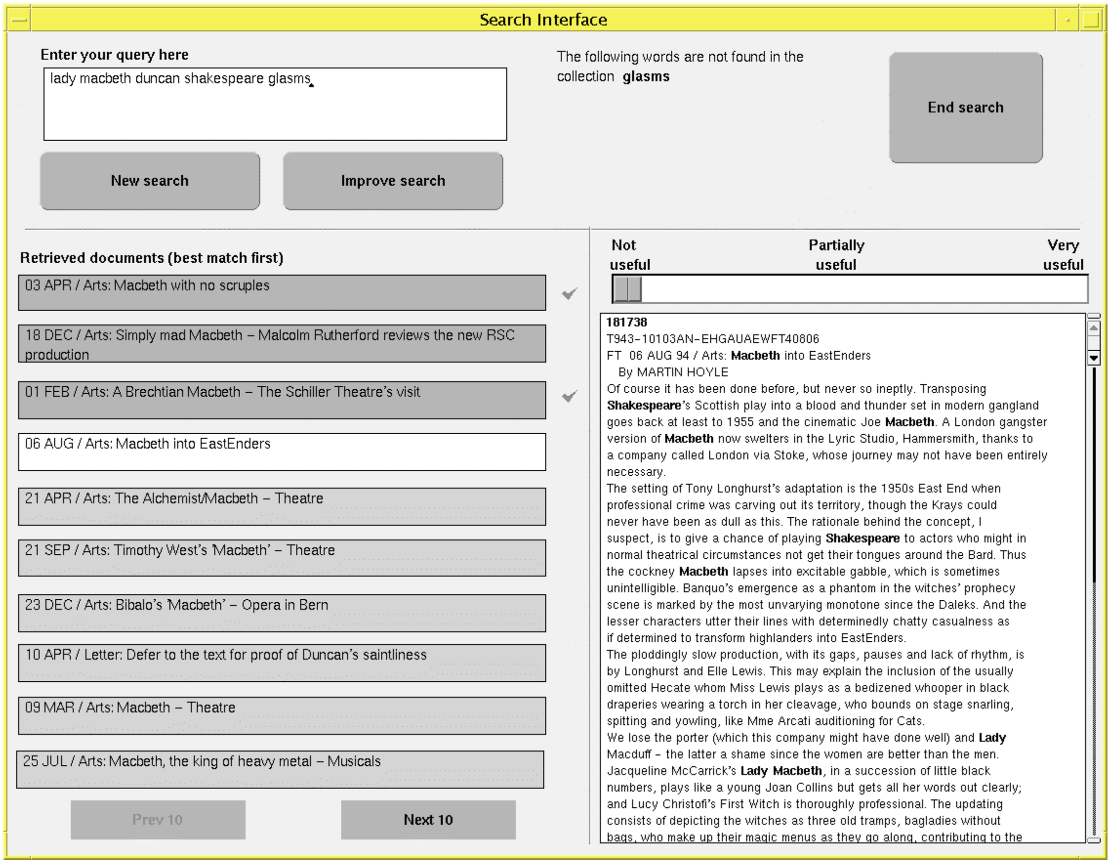
Simulated situation 307i
The new Scottish Parliament is considering planning permission for a series of large hydroelectric projects. These projects will use water power to produce electricity for a large area of Scotland. Supporters of the projects claim that they will give cheaper electricity and reduce 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.

Simulated situation 321
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.

Simulated situation 322i
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.

Simulated situation 326i
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.

Simulated situation 347i
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.



lady macbeth duncan shakespeare glasms

The following words are not found in the collection **glasms**

- 03 APR / Arts: Macbeth with no scruples ✓
- 18 DEC / Arts: Simply mad Macbeth – Malcolm Rutherford reviews the new RSC production
- 01 FEB / Arts: A Brechtian Macbeth – The Schiller Theatre's visit ✓
- 06 AUG / Arts: Macbeth into EastEnders
- 21 APR / Arts: The Alchemist/Macbeth – Theatre
- 21 SEP / Arts: Timothy West's 'Macbeth' – Theatre
- 23 DEC / Arts: Bibalo's 'Macbeth' – Opera in Bern
- 10 APR / Letter: Defer to the text for proof of Duncan's saintliness
- 09 MAR / Arts: Macbeth – Theatre
- 25 JUL / Arts: Macbeth, the king of heavy metal – Musicals

181738
T943-10103AN-EHGAUAEWFT40806
FT 06 AUG 94 / Arts: **Macbeth** into EastEnders
By MARTIN HOYLE

Of course it has been done before, but never so ineptly. Transposing **Shakespeare**'s Scottish play into a blood and thunder set in modern gangland goes back at least to 1955 and the cinematic Joe **Macbeth**. A London gangster version of **Macbeth** now swelters in the Lyric Studio, Hammersmith, thanks to a company called London via Stoke, whose journey may not have been entirely necessary.

The setting of Tony Longhurst's adaptation is the 1950s East End when professional crime was carving out its territory, though the Krays could never have been as dull as this. The rationale behind the concept, I suspect, is to give a chance of playing **Shakespeare** to actors who might in normal theatrical circumstances not get their tongues around the Bard. Thus the cockney **Macbeth** lapses into excitable gabble, which is sometimes unintelligible. Banquo's emergence as a phantom in the witches' prophecy scene is marked by the most unvarying monotone since the Daleks. And the lesser characters utter their lines with determinedly chatty casualness as if determined to transform highlanders into EastEnders.

The ploddingly slow production, with its gaps, pauses and lack of rhythm, is by Longhurst and Elle Lewis. This may explain the inclusion of the usually omitted Hecate whom Miss Lewis plays as a bedizened whooper in black draperies wearing a torch in her cleavage, who bounds on stage snarling, spitting and yowling, like Mme Arcati auditioning for Cats.

We lose the porter (which this company might have done well) and **Lady** Macduff – the latter a shame since the women are better than the men. Jacqueline McCarrick's **Lady Macbeth**, in a succession of little black numbers, plays like a young Joan Collins but gets all her words out clearly; and Lucy Christofi's First Witch is thoroughly professional. The updating consists of depicting the witches as three old tramps, bagladies without bags, who make up their magic menus as they go along, contributing to the

Figure A.1: Search interface for Experiment Two