A Study of Factors Affecting the Utility of Implicit Relevance Feedback

Ryen W. White
Human-Computer Interaction Laboratory
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
ryen@umd.edu

Ian Ruthven
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland. G1 1XH.
ir@cis.strath.ac.uk

Joemon M. Jose
Department of Computing Science
University of Glasgow
Glasgow, Scotland. G12 8RZ.
jj@dcs.gla.ac.uk

ABSTRACT
Implicit relevance feedback (IRF) is the process by which a search system unobtrusively gathers evidence on searcher interests from their interaction with the system. IRF is a new method of gathering information on user interest and, if IRF is to be used in operational IR systems, it is important to establish when it performs well and when it performs poorly. In this paper we investigate how the use and effectiveness of IRF is affected by three factors: search task complexity, the search experience of the user and the stage in the search. Our findings suggest that all three of these factors contribute to the utility of IRF.

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

General Terms
Experimentation, Human Factors.

Keywords
Implicit Relevance Feedback, Relevance Feedback

1. INTRODUCTION
Information Retrieval (IR) systems are designed to help searchers solve problems. In the traditional interaction metaphor employed by Web search systems such as Yahoo! and MSN Search, the system generally only supports the retrieval of potentially relevant documents from the collection. However, it is also possible to offer support to searchers for different search activities, such as selecting the terms to present to the system or choosing which search strategy to adopt [3, 8]; both of which can be problematic for searchers.

As the quality of the query submitted to the system directly affects the quality of search results, the issue of how to improve search queries has been studied extensively in IR research [6]. Techniques such as Relevance Feedback (RF) [11] have been proposed as a way in which the IR system can support the iterative development of a search query by suggesting alternative terms for query modification. However, in practice RF techniques have been underutilised as they place an increased cognitive burden on searchers to directly indicate relevant results [10].

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

Implicit Relevance Feedback (IRF) [7] has been proposed as a way in which search queries can be improved by passively observing searchers as they interact. IRF has been implemented either through the use of surrogate measures based on interaction with documents (such as reading time, scrolling or document retention) [7] or using interaction with browse-based result interfaces [5]. IRF has been shown to display mixed effectiveness because the factors that are good indicators of user interest are often erratic and the inferences drawn from user interaction are not always valid [7].

In this paper we present a study into the use and effectiveness of IRF in an online search environment. The study aims to investigate the factors that affect IRF, in particular three research questions: (i) is the use of and perceived quality of terms generated by IRF affected by the search task? (ii) is the use of and perceived quality of terms generated by IRF affected by the level of search experience of system users? (iii) is IRF equally used and does it generate terms that are equally useful at all search stages? This study aims to establish when, and under what circumstances, IRF performs well in terms of its use and the query modification terms selected as a result of its use.

The main experiment from which the data are taken was designed to test techniques for selecting query modification terms and techniques for displaying retrieval results [13]. In this paper we use data derived from that experiment to study factors affecting the utility of IRF.

2. STUDY
In this section we describe the user study conducted to address our research questions.

2.1 Systems
Our study used two systems both of which suggested new query terms to the user. One system suggested terms based on the user's interaction (IRF), the other used Explicit RF (ERF) asking the user to explicitly indicate relevant material. Both systems used the same term suggestion algorithm, [15], and used a common interface.

2.1.1 Interface Overview
In both systems, retrieved documents are represented at the interface by their full-text and a variety of smaller, query-relevant representations, created at retrieval time. We used the Web as the test collection in this study and Google[1] as the underlying search engine. Document representations include the document title and a summary of the document; a list of *top-ranking sentences* (TRS) extracted from the top documents retrieved, scored in relation to the query, a sentence in the document summary, and each summary sentence in the context

[1] http://www.google.com/

it occurs in the document (i.e., with the preceding and following sentence). Each summary sentence and top-ranking sentence is regarded as a representation of the document. The default display contains the list of top-ranking sentences and the list of the first ten document titles. Interacting with a representation guides searchers to a different representation from the same document, e.g., moving the mouse over a document title displays a summary of the document. This presentation of progressively more information from documents to aid relevance assessments has been shown to be effective in earlier work [14, 16]. In Appendix A we show the complete interface to the IRF system with the document representations marked and in Appendix B we show a fragment from the ERF interface with the checkboxes used by searchers to indicate relevant information. Both systems provide an interactive query expansion feature by suggesting new query terms to the user. The searcher has the responsibility for choosing which, if any, of these terms to add to the query. The searcher can also add or remove terms from the query at will.

2.1.2 Explicit RF system

This version of the system implements explicit RF. Next to each document representation are checkboxes that allow searchers to mark individual representations as relevant; *marking* a representation is an indication that its contents are relevant. Only the representations marked relevant by the user are used for suggesting new query terms. This system was used as a baseline against which the IRF system could be compared.

2.1.3 Implicit RF system

This system makes inferences about searcher interests based on the information with which they interact. As described in Section 2.1.1 interacting with a representation highlights a new representation from the same document. To the searcher this is a way they can find out more information from a potentially interesting source. To the implicit RF system each interaction with a representation is interpreted as an implicit indication of interest in that representation; *interacting with* a representation is assumed to be an indication that its contents are relevant. The query modification terms are selected using the same algorithm as in the Explicit RF system. Therefore the only difference between the systems is how relevance is communicated to the system.

The results of the main experiment [13] indicated that these two systems were comparable in terms of effectiveness.

2.2 Tasks

Search tasks were designed to encourage realistic search behaviour by our subjects. The tasks were phrased in the form of simulated work task situations [2], i.e., short search scenarios that were designed to reflect real-life search situations and allow subjects to develop personal assessments of relevance. We devised six search topics (i.e., applying to university, allergies in the workplace, art galleries in Rome, "Third Generation" mobile phones, Internet music piracy and petrol prices) based on pilot testing with a small representative group of subjects. These subjects were not involved in the main experiment.

For each topic, three versions of each work task situation were devised, each version differing in their predicted level of task *complexity*. As described in [1] task complexity is a variable that affects subject perceptions of a task and their interactive behaviour, e.g., subjects perform more filtering activities with highly complex search tasks. By developing tasks of different complexity we can assess how the nature of the task affects the subjects' interactive behaviour and hence the evidence supplied to IRF algorithms. Task complexity was varied according to the methodology described in [1], specifically by varying the number of potential information sources and types of information required, to complete a task. In our pilot tests (and in *a posteriori* analysis of the main experiment results) we verified that subjects reporting of individual task complexity matched our estimation of the complexity of the task.

Subjects attempted three search tasks: one high complexity, one moderate complexity and one low complexity[2]. They were asked to read the task, place themselves in the situation it described and find the information they felt was required to complete the task. Figure 1 shows the task statements for three levels of task complexity for one of the six search topics.

HC Task: High Complexity
Whilst having dinner with an American colleague, they comment on the high price of petrol in the UK compared to other countries, despite large volumes coming from the same source. Unaware of any major differences, you decide to find out how and why petrol prices vary worldwide.

MC Task: Moderate Complexity
Whilst out for dinner one night, one of your friends' guests is complaining about the price of petrol and the factors that cause it. Throughout the night they seem to be complaining about everything they can, reducing the credibility of their earlier statements so you decide to research which factors actually are important in determining the price of petrol in the UK.

LC Task: Low Complexity
While out for dinner one night, your friend complains about the rising price of petrol. However, as you have not been driving for long, you are unaware of any major changes in price. You decide to find out how the price of petrol has changed in the UK in recent years.

Figure 1. Varying task complexity ("Petrol Prices" topic).

2.3 Subjects

156 volunteers expressed an interest in participating in our study. 48 subjects were selected from this set with the aim of populating two groups, each with 24 subjects: *inexperienced* (infrequent/ inexperienced searchers) and *experienced* (frequent/ experienced searchers). Subjects were not chosen and classified into their groups until they had completed an entry questionnaire that asked them about their search experience and computer use.

The average age of the subjects was 22.83 years (maximum 51, minimum 18, σ = 5.23 years) and 75% had a university diploma or a higher degree. 47.91% of subjects had, or were pursuing, a qualification in a discipline related to Computer Science. The subjects were a mixture of students, researchers, academic staff and others, with different levels of computer and search experience. The subjects were divided into the two groups depending on their search experience, how often they searched and the types of searches they performed. All were familiar with Web searching, and some with searching in other domains.

2.4 Methodology

The experiment had a factorial design; with 2 levels of search experience, 3 experimental systems (although we only report on the findings from the ERF and IRF systems) and 3 levels of search task complexity. Subjects attempted one task of each complexity,

[2] The main experiment from which these results are drawn had a third comparator system which had a different interface. Each subject carried out three tasks, one on each system. We only report on the results from the ERF and IRF systems as these are the only pertinent ones for this paper.

switched systems after each task and used each system once. The order in which systems were used and search tasks attempted was randomised according to a Latin square experimental design. Questionnaires used Likert scales, semantic differentials and open-ended questions to elicit subject opinions [4]. System logging was also used to record subject interaction.

A tutorial carried out prior to the experiment allowed subjects to use a non-feedback version of the system to attempt a practice task before using the first experimental system. Experiments lasted between one-and-a-half and two hours, dependent on variables such as the time spent completing questionnaires. Subjects were offered a 5 minute break after the first hour. In each experiment:

i. the subject was welcomed and asked to read an introduction to the experiments and sign consent forms. This set of instructions was written to ensure that each subject received precisely the same information.
ii. the subject was asked to complete an introductory questionnaire. This contained questions about the subject's education, general search experience, computer experience and Web search experience.
iii. the subject was given a tutorial on the interface, followed by a training topic on a version of the interface with no RF.
iv. the subject was given three task sheets and asked to choose one task from the six topics on each sheet. No guidelines were given to subjects when choosing a task other than they could not choose a task from any topic more than once. Task complexity was rotated by the experimenter so each subject attempted one high complexity task, one moderate complexity task and one low complexity task.
v. the subject was asked to perform the search and was given 15 minutes to search. The subject could terminate a search early if they were unable to find any more information they felt helped them complete the task.
vi. after completion of the search, the subject was asked to complete a post-search questionnaire.
vii. the remaining tasks were attempted by the subject, following steps v. and vi.
viii. the subject completed a post-experiment questionnaire and participated in a post-experiment interview.

Subjects were told that their interaction may be used by the IRF system to help them as they searched. They were not told which behaviours would be used or how it would be used.

We now describe the findings of our analysis.

3. FINDINGS

In this section we use the data derived from the experiment to answer our research questions about the effect of search task complexity, search experience and stage in search on the use and effectiveness of IRF. We present our findings per research question. Due to the ordinal nature of much of the data non-parametric statistical testing is used in this analysis and the level of significance is set to $p < .05$, unless otherwise stated. We use the method proposed by [12] to determine the significance of differences in multiple comparisons and that of [9] to test for interaction effects between experimental variables, the occurrence of which we report where appropriate. All Likert scales and semantic differentials were on a 5-point scale where a rating closer to 1 signifies more agreement with the attitude statement. The category labels HC, MC and LC are used to denote the high, moderate and low complexity tasks respectively. The highest, or most positive, values in each table are shown in bold. Our analysis uses data from questionnaires, post-experiment interviews and background system logging on the ERF and IRF systems.

3.1 Search Task

Searchers attempted three search tasks of varying complexity, each on a different experimental system. In this section we present an analysis on the use and usefulness of IRF for search tasks of different complexities. We present our findings in terms of the RF provided by subjects and the terms recommended by the systems.

3.1.1 Feedback

We use questionnaires and system logs to gather data on subject perceptions and provision of RF for different search tasks. In the post-search questionnaire subjects were asked about how RF was conveyed using differentials to elicit their opinion on:

1. the *value* of the feedback technique: *How you conveyed relevance to the system (i.e. ticking boxes or viewing information) was*: "easy" / "difficult", "effective"/ "ineffective", "useful'"/"not useful".

2. the *process* of providing the feedback: *How you conveyed relevance to the system made you feel:* "comfortable"/"uncomfortable", "in control"/"not in control".

The average obtained differential values are shown in Table 1 for IRF and each task category. The value corresponding to the differential "*All*" represents the mean of all differentials for a particular attitude statement. This gives some overall understanding of the subjects' feelings which can be useful as the subjects may not answer individual differentials very precisely. The values for ERF are included for reference in this table and all other tables and figures in the "Findings" section. Since the aim of the paper is to investigate situations in which IRF might perform well, not a direct comparison between IRF and ERF, we make only limited comparisons between these two types of feedback.

Table 1. Subject perceptions of RF method (lower = better).

Differential	Explicit RF			Implicit RF		
	HC	MC	LC	HC	MC	LC
Easy	2.78	2.47	**2.12**	1.86	**1.81**	1.93
Effective	2.94	2.68	**2.44**	**2.04**	2.41	2.66
Useful	2.76	2.51	**2.16**	**1.91**	2.37	2.56
All (1)	2.83	2.55	**2.24**	**1.94**	2.20	2.38
Comfortable	**2.27**	2.28	2.35	**2.11**	2.15	2.16
In control	2.01	**1.97**	1.93	2.73	2.68	**2.61**
All (2)	2.14	**2.13**	2.14	2.42	2.42	**2.39**

Each cell in Table 1 summarises the subject responses for 16 task-system pairs (16 subjects who ran a high complexity (HC) task on the ERF system, 16 subjects who ran a medium complexity (MC) task on the ERF system, etc). Kruskal-Wallis Tests were applied to each differential for each type of RF[3]. Subject responses suggested that

[3] Since this analysis involved many differentials, we use a Bonferroni correction to control the experiment-wise error rate and set the *alpha level* (α) to .0167 and .0250 for both statements 1. and 2. respectively, i.e., .05 divided by the number of differentials. This correction reduces the number of Type I errors i.e., rejecting null hypotheses that are true.

IRF was most "*effective*" and "*useful*" for more complex search tasks[4] and that the differences in all pair-wise comparisons between tasks were significant[5]. Subject perceptions of IRF elicited using the other differentials did not appear to be affected by the complexity of the search task[6]. To determine whether a relationship exists between the effectiveness and usefulness of the IRF process and task complexity we applied Spearman's Rank Order Correlation Coefficient to participant responses. The results of this analysis suggest that the effectiveness of IRF and usefulness of IRF are both related to task complexity; as task complexity increases subject preference for IRF also increases[7].

On the other hand, subjects felt ERF was more "*effective*" and "*useful*" for low complexity tasks[8]. Their verbal reporting of ERF, where perceived utility and effectiveness increased as task complexity decreased, supports this finding. In tasks of lower complexity the subjects felt they were better able to provide feedback on whether or not documents were relevant to the task.

We analyse interaction logs generated by both interfaces to investigate the amount of RF subjects provided. To do this we use a measure of search "precision" that is the proportion of all possible document representations that a searcher assessed, divided by the total number they could assess. In ERF this is the proportion of all possible representations that were marked relevant by the searcher, i.e., those representations explicitly marked relevant. In IRF this is the proportion of representations viewed by a searcher over all possible representations that *could* have been viewed by the searcher. This proportion measures the searcher's level of interaction with a document, we take it to measure the user's interest in the document: the more document representations viewed the more interested we assume a user is in the content of the document.

There are a maximum of 14 representations per document: 4 top-ranking sentences, 1 title, 1 summary, 4 summary sentences and 4 summary sentences in document context. Since the interface shows document representations from the top-30 documents, there are 420 representations that a searcher can assess. Table 2 shows proportion of representations provided as RF by subjects.

Table 2. Feedback and documents viewed.

Measure	Explicit RF			Implicit RF		
	HC	MC	LC	HC	MC	LC
Proportion Feedback	2.14	2.39	**2.65**	**21.50**	19.36	15.32
Documents Viewed	10.63	10.43	**10.81**	10.84	12.19	**14.81**

For IRF there is a clear pattern: as complexity increases the subjects viewed fewer documents but viewed more representations for each document. This suggests a pattern where users are investigating retrieved documents in more depth. It also means that the amount of feedback varies based on the complexity of the search task. Since IRF is based on the interaction of the searcher, the more they interact, the more feedback they provide. This has no effect on the number of RF terms chosen, but may affect the quality of the terms selected.

Correlation analysis revealed a strong negative correlation between the number of documents viewed and the amount of feedback searchers provide[9]; as the number of documents viewed increases the proportion of feedback falls (searchers view less representations of each document). This may be a natural consequence of their being less time to view documents in a time constrained task environment but as we will show as complexity changes, the nature of information searchers interact with also appears to change. In the next section we investigate the effect of task complexity on the terms chosen as a result of IRF.

3.1.2 Terms

The same RF algorithm was used to select query modification terms in all systems [16]. We use subject opinions of terms recommended by the systems as a measure of the effectiveness of IRF with respect to the terms generated for different search tasks. To test this, subjects were asked to complete two semantic differentials that completed the statement: *The words chosen by the system were*: "relevant"/"irrelevant" and "useful"/"not useful". Table 3 presents average responses grouped by search task.

Table 3. Subject perceptions of system terms (lower = better).

Differential	Explicit RF			Implicit RF		
	HC	MC	LC	HC	MC	LC
Relevant	2.50	2.46	**2.41**	**1.94**	2.35	2.68
Useful	2.61	2.61	**2.59**	**2.06**	2.54	2.70

Kruskal-Wallis Tests were applied within each type of RF. The results indicate that the relevance and usefulness of the terms chosen by IRF is affected by the complexity of the search task; the terms chosen are more "*relevant*" and "*useful*" when the search task is more complex. [10] Relevant here, was explained as being related to their task whereas useful was for terms that were seen as being helpful in the search task. For ERF, the results indicate that the terms generated are perceived to be more "*relevant*" and "*useful"* for less complex search tasks; although differences between tasks were not significant[11]. This suggests that subject perceptions of the terms chosen for query modification are affected by task complexity. Comparison between ERF and IRF shows that subject perceptions also vary for different types of RF[12].

As well as using data on relevance and utility of the terms chosen, we used data on term *acceptance* to measure the perceived value of the terms suggested. Explicit and Implicit RF systems made recommendations about which terms could be added to the original search query. In Table 4 we show the proportion of the top six terms

[4] *effective*: $\chi^2(2) = 11.62$, p = .003; *useful*: $\chi^2(2) = 12.43$, p = .002

[5] Dunn's *post-hoc* tests (multiple comparison using rank sums); all $Z \geq 2.88$, all $p \leq .002$

[6] all $\chi^2(2) \leq 2.85$, all $p \geq .24$ (Kruskal-Wallis Tests)

[7] *effective*: all $r \geq 0.644$, $p \leq .002$; *useful*: all $r \geq 0.541$, $p \leq .009$

[8] *effective*: $\chi^2(2) = 7.01$, p = .03; *useful*: $\chi^2(2) = 6.59$, p = .037 (Kruskal-Wallis Test); all pair-wise differences significant, all $Z \geq 2.34$, all $p \leq .01$ (Dunn's *post-hoc* tests)

[9] $r = -0.696$, p = .001 (Pearson's Correlation Coefficient)

[10] *relevant*: $\chi^2(2) = 13.82$, p = .001; *useful*: $\chi^2(2) = 11.04$, p = .004; $\alpha = .025$

[11] all $\chi^2(2) \leq 2.28$, all $p \geq .32$ (Kruskal-Wallis Test)

[12] all $T(16) \geq 102$, all $p \leq .021$, (Wilcoxon Signed-Rank Test)

[13] that were shown to the searcher that were added to the search query, for each type of task and each type of RF.

Table 4. Term Acceptance (percentage of top six terms).

Proportion of terms	Explicit RF			Implicit RF		
	HC	MC	LC	HC	MC	LC
Accepted	65.31	67.32	**68.65**	67.45	67.24	**67.59**

The average number of terms accepted from IRF is approximately the same across all search tasks and generally the same as that of ERF[14]. As Table 2 shows, subjects marked fewer documents relevant for highly complex tasks . Therefore, when task complexity increases the ERF system has fewer examples of relevant documents and the expansion terms generated may be poorer. This could explain the difference in the proportion of recommended terms accepted in ERF as task complexity increases. For IRF there is little difference in how many of the recommended terms were chosen by subjects for each level of task complexity[15]. Subjects may have perceived IRF terms as more useful for high complexity tasks but this was not reflected in the proportion of IRF terms accepted. Differences may reside in the nature of the terms accepted; future work will investigate this issue.

3.1.3 Summary

In this section we have presented an investigation on the effect of search task complexity on the utility of IRF. From the results there appears to be a strong relation between the complexity of the task and the subject interaction: subjects preferring IRF for highly complex tasks. Task complexity did not affect the proportion of terms accepted in either RF method, despite there being a difference in how "*relevant*" and "*useful*" subjects perceived the terms to be for different complexities; complexity may affect term selection in ways other than the proportion of terms accepted.

3.2 Search Experience

Experienced searchers may interact differently and give different types of evidence to RF than inexperienced searchers. As such, levels of search experience may affect searchers' use and perceptions of IRF. In our experiment subjects were divided into two groups based on their level of search experience, the frequency with which they searched and the types of searches they performed. In this section we use their perceptions and logging to address the next research question; the relationship between the usefulness and use of IRF and the search experience of experimental subjects. The data are the same as that analysed in the previous section, but here we focus on search experience rather than the search task.

3.2.1 Feedback

We analyse the results from the attitude statements described at the beginning of Section 3.1.1. (i.e., *How you conveyed relevance to the system was…* and *How you conveyed relevance to the system made you feel…*). These differentials elicited opinion from experimental subjects about the RF method used. In Table 5 we show the mean average responses for inexperienced and experienced subject groups on ERF and IRF; 24 subjects per cell.

[13] This was the smallest number of query modification terms that were offered in both systems.

[14] all $T(16) \geq 80$, all $p \leq .31$, (Wilcoxon Signed-Rank Test)

[15] ERF: $\chi^2(2) = 3.67$, $p = .16$; IRF: $\chi^2(2) = 2.55$, $p = .28$ (Kruskal-Wallis Tests)

Table 5. Subject perceptions of RF method (lower = better).

Differential	Explicit RF		Implicit RF	
	Inexp.	Exp.	Inexp.	Exp.
Easy	**2.46**	**2.46**	**1.84**	1.98
Effective	2.75	**2.63**	**2.32**	2.43
Useful	2.50	**2.46**	2.28	**2.27**
All (1)	2.57	**2.52**	**2.14**	2.23
Comfortable	2.46	**2.14**	**2.05**	2.24
In control	**1.96**	1.98	2.73	**2.64**
All (2)	2.21	**2.06**	**2.39**	2.44

The results demonstrate a strong preference in inexperienced subjects for IRF; they found it more "*easy*" and "*effective*" than experienced subjects. [16] The differences for all other IRF differentials were not statistically significant. For all differentials, apart from "*in control*", inexperienced subjects generally preferred IRF over ERF[17]. Inexperienced subjects also felt that IRF was more difficult to control than experienced subjects[18]. As these subjects have less search experience they may be less able to understand RF processes and may be more comfortable with the system gathering feedback implicitly from their interaction. Experienced subjects tended to like ERF more than inexperienced subjects and felt more "*comfortable*" with this feedback method[19]. It appears from these results that experienced subjects found ERF more useful and were more at ease with the ERF process.

In a similar way to Section 3.1.1 we analysed the proportion of feedback that searchers provided to the experimental systems. Our analysis suggested that search experience does not affect the amount of feedback subjects provide[20].

3.2.2 Terms

We used questionnaire responses to gauge subject opinion on the relevance and usefulness of the terms from the perspective of experienced and inexperienced subjects. Table 6 shows the average differential responses obtained from both subject groups.

Table 6. Subject perceptions of system terms (lower = better).

Differential	Explicit RF		Implicit RF	
	Inexp.	Exp.	Inexp.	Exp.
Relevant	2.58	**2.44**	2.33	**2.21**
Useful	2.88	**2.63**	2.33	**2.23**

The differences between subject groups were significant[21]. Experienced subjects generally reacted to the query modification terms chosen by the system more positively than inexperienced

[16] *easy*: $U(24) = 391$, $p = .016$; *effective*: $U(24) = 399$, $p = .011$; $\alpha = .0167$ (Mann-Whitney Tests)

[17] all $T(24) \geq 231$, all $p \leq .001$ (Wilcoxon Signed-Rank Test)

[18] $U(24) = 390$, $p = .018$; $\alpha = .0250$ (Mann-Whitney Test)

[19] $T(24) = 222$, $p = .020$ (Wilcoxon Signed-Rank Test)

[20] ERF: all $U(24) \leq 319$, $p \geq .26$, IRF: all $U(24) \leq 313$, $p \geq .30$ (Mann-Whitney Tests)

[21] ERF: all $U(24) \geq 388$, $p \leq .020$, IRF: all $U(24) \geq 384$, $p \leq .024$

subjects. This finding was supported by the proportion of query modification terms these subjects accepted. In the same way as in Section 3.1.2, we analysed the number of query modification terms recommended by the system that were used by experimental subjects. Table 7 shows the average number of accepted terms per subject group.

Table 7. Term Acceptance (percentage of top six terms).

Proportion of terms	Explicit RF		Implicit RF	
	Inexp.	Exp.	Inexp.	Exp.
Accepted	63.76	**70.44**	64.43	**71.35**

Our analysis of the data show that differences between subject groups for each type of RF are significant; experienced subjects accepted more expansion terms regardless of type of RF. However, the differences between the same groups for different types of RF are not significant; subjects chose roughly the same percentage of expansion terms offered irrespective of the type of RF[22].

3.2.3 Summary

In this section we have analysed data gathered from two subject groups – inexperienced searchers and experienced searchers – on how they perceive and use IRF. The results indicate that inexperienced subjects found IRF more "*easy*" and "*effective*" than experienced subjects, who in turn found the terms chosen as a result of IRF more "*relevant*" and "*useful*". We also showed that inexperienced subjects generally accepted less recommended terms than experienced subjects, perhaps because they were less comfortable with RF or generally submitted shorter search queries. Search experience appears to affect how subjects use the terms recommended as a result of the RF process.

3.3 Search Stage

From our observations of experimental subjects as they searched we conjectured that RF may be used differently at different times during a search. To test this, our third research question concerned the use and usefulness of IRF during the course of a search. In this section we investigate whether the amount of RF provided by searchers or the proportion of terms accepted are affected by how far through their search they are. For the purposes of this analysis a search begins when a subject poses the first query to the system and progresses until they terminate the search or reach the maximum allowed time for a search task of 15 minutes. We do not divide tasks based on this limit as subjects often terminated their search in less than 15 minutes.

In this section we use data gathered from interaction logs and subject opinions to investigate the extent to which RF was used and the extent to which it appeared to benefit our experimental subjects at different stages in their search

3.3.1 Feedback

The interaction logs for all searches on the Explicit RF and Implicit RF were analysed and each search is divided up into nine equal length time slices. This number of slices gave us an equal number per stage and was a sufficient level of granularity to identify trends in the results. Slices 1 – 3 correspond to the "start" of the search, 4 – 6 to the "middle" of the search and 7 – 9 to the "end". In Figure 2 we plot the measure of "precision" described in Section 3.1.1 (i.e., the proportion of all possible representations that were provided as RF) at each of the nine slices, per search task, averaged across all subjects; this allows us to see how the provision of RF was distributed during a search. The total amount of feedback for a single RF method/task complexity pairing across all nine slices corresponds to the value recorded in the first row of Table 2 (e.g., the sum of the RF for IRF/HC across all nine slices of Figure 2 is 21.50%). To simplify the statistical analysis and comparison we use the grouping of "start", "middle" and "end".

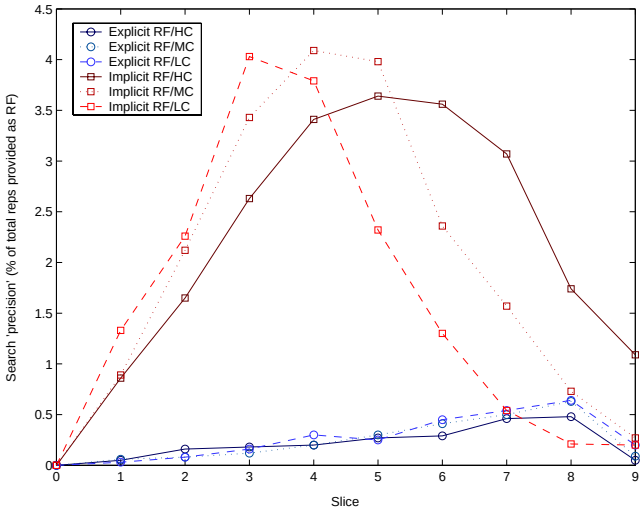


Figure 2. Distribution of RF provision per search task.

Figure 2 appears to show the existence of a relationship between the stage in the search and the amount of relevance information provided to the different types of feedback algorithm. These are essentially differences in the way users are assessing documents. In the case of ERF subjects provide explicit relevance assessments throughout most of the search, but there is generally a steep increase in the "end" phase towards the completion of the search[23].

When using the IRF system, the data indicates that at the start of the search subjects are providing little relevance information[24], which corresponds to interacting with few document representations. At this stage the subjects are perhaps concentrating more on reading the retrieved results. Implicit relevance information is generally offered extensively in the middle of the search as they interact with results and it then tails off towards the end of the search. This would appear to correspond to stages of initial exploration, detailed analysis of document representations and storage and presentation of findings.

Figure 2 also shows the proportion of feedback for tasks of different complexity. The results appear to show a difference[25] in how IRF is used that relates to the complexity of the search task. More specifically, as complexity increases it appears as though subjects take longer to reach their most interactive point. This suggests that task complexity affects how IRF is distributed during the search and that they may be spending more time initially interpreting search results for more complex tasks.

[22] IRF: $U(24) = 403$, p = .009, ERF: $U(24) = 396$, p = .013

[23] IRF: all $Z \geq 1.87$, $p \leq .031$, ERF: "start" vs. "end" $Z = 2.58$, p = .005 (Dunn's *post-hoc* tests).

[24] Although increasing toward the end of the "start" stage.

[25] Although not statistically significant; $\chi^2(2) = 3.54$, p = .17 (Friedman Rank Sum Test)

3.3.2 Terms

The terms recommended by the system are chosen based on the frequency of their occurrence in the relevant items. That is, non-stopword, non-query terms occurring frequently in search results regarded as relevant are likely to be recommended to the searcher for query modification. Since there is a direct association between the RF and the terms selected we use the number of terms accepted by searchers at different points in the search as an indication of how effective the RF has been up until the current point in the search. In this section we analysed the average number of terms from the top six terms recommended by Explicit RF and Implicit RF over the course of a search. The average proportion of the top six recommended terms that were accepted at each stage are shown in Table 8; each cell contains data from all 48 subjects.

Table 8. Term Acceptance (proportion of top six terms).

Proportion of terms	Explicit RF			Implicit RF		
	start	middle	end	start	middle	end
Accepted	66.87	66.98	**67.34**	61.85	68.54	**73.22**

The results show an apparent association between the stage in the search and the number of feedback terms subjects accept. Search stage affects term acceptance in IRF but not in ERF[26]. The further into a search a searcher progresses, the more likely they are to accept terms recommended via IRF (significantly more than ERF[27]). A correlation analysis between the proportion of terms accepted at each search stage and *cumulative* RF (i.e., the sum of all "precision" at each slice in Figure 2 up to and including the end of the search stage) suggests that in both types of RF the quality of system terms improves as more RF is provided[28].

3.3.3 Summary

The results from this section indicate that the location in a search affects the amount of feedback given by the user to the system, and hence the amount of information that the RF mechanism has to decide which terms to offer the user. Further, trends in the data suggest that the complexity of the task affects how subjects provide IRF and the proportion of system terms accepted.

4. DISCUSSION AND IMPLICATIONS

In this section we discuss the implications of the findings presented in the previous section for each research question.

4.1 Search Task

The results of our study showed that ERF was preferred for less complex tasks and IRF for more complex tasks. From observations and subject comments we perceived that when using ERF systems subjects generally forgot to provide the feedback but also employed different criteria during the ERF process (i.e., they were assessing relevance rather than expressing an interest). When the search was more complex subjects rarely found results they regarded as completely relevant. Therefore they struggled to find relevant information and were unable to communicate RF to the search system. In these situations subjects appeared to prefer IRF as they do not need to make a relevance decision to obtain the benefits of RF, i.e., term suggestions, whereas in ERF they do.

The association between RF method and task complexity has implications for the design of user studies of RF systems and the RF systems themselves. It implies that in the design of user studies involving ERF or IRF systems care should be taken to include tasks of varying complexities, to avoid task bias. Also, in the design of search systems it implies that since different types of RF may be appropriate for different task complexities then a system that could automatically detect complexity could use both ERF and IRF simultaneously to benefit the searcher. For example, on the IRF system we noticed that as task complexity falls search behaviour shifts from results interface to retrieved documents. Monitoring such interaction across a number of studies may lead to a set of criteria that could help IR systems automatically detect task complexity and tailor support to suit.

4.2 Search Experience

We analysed the affect of search experience on the utility of IRF. Our analysis revealed a general preference across all subjects for IRF over ERF. That is, the average ratings assigned to IRF were generally more positive than those assigned to ERF. However, IRF was generally liked by both subject groups (perhaps because it removed the burden of providing relevance information) and ERF was generally preferred by experienced subjects more than inexperienced subjects (perhaps because it allowed them to specify which results were used by the system when generating term recommendations).

All subjects felt more in control with ERF than IRF, but for inexperienced subjects this did not appear to affect their overall preferences[29]. These subjects may understand the RF process less, but may be more willing to sacrifice control over feedback in favour of IRF, a process that they perceive more positively.

4.3 Search Stage

We also analysed the effects of search stage on the use and usefulness of IRF. Through analysis of this nature we can build a more complete picture of how searchers used RF and how this varies based on the RF method. The results suggest that IRF is used more in the middle of the search than at the beginning or end, whereas ERF is used more towards the end. The results also show the effects of task complexity on the IRF process and how rapidly subjects reach their most interactive point. Without an analysis of this type it would not have been possible to establish the existence of such patterns of behaviour.

The findings suggest that searchers interact differently for IRF and ERF. Since ERF is not traditionally used until toward the end of the search it may be possible to incorporate both IRF and ERF into the same IR system, with IRF being used to gather evidence until subjects decide to use ERF. The development of such a system represents part of our ongoing work in this area.

5. CONCLUSIONS

In this paper we have presented an investigation of Implicit Relevance Feedback (IRF). We aimed to answer three research questions about factors that may affect the provision and usefulness of IRF. These factors were search task complexity, the subjects' search experience and the stage in the search. Our overall conclusion was that all factors

[26] ERF: $\chi^2(2) = 2.22$, $p = .33$; IRF: $\chi^2(2) = 7.73$, $p = .021$ (Friedman Rank Sum Tests); IRF: all pair-wise comparisons significant at $Z \geq 1.77$, all $p \leq .038$ (Dunn's *post-hoc* tests)

[27] all $T(48) \geq 786$, all $p \leq .002$, (Wilcoxon Signed-Rank Test)

[28] IRF: $r = .712$, $p < .001$, ERF: $r = .695$, $p = .001$ (Pearson Correlation Coefficient)

[29] This may also be true for experienced subjects, but the data we have is insufficient to draw this conclusion.

appear to have some effect on the use and effectiveness of IRF, although the interaction effects between factors are not statistically significant.

Our conclusions per each research question are: (i) IRF is generally more useful for complex search tasks, where searchers want to focus on the search task and get new ideas for their search from the system, (ii) IRF is preferred to ERF overall and generally preferred by inexperienced subjects wanting to reduce the burden of providing RF, and (iii) within a single search session IRF is affected by temporal location in a search (i.e., it is used in the middle, not the beginning or end) and task complexity.

Studies of this nature are important to establish the circumstances where a promising technique such as IRF are useful and those when it is not. It is only after such studies have been run and analysed in this way can we develop an understanding of IRF that allow it to be successfully implemented in operational IR systems.

6. REFERENCES

[1] Bell, D.J. and Ruthven, I. (2004). Searchers' assessments of task complexity for web searching. *Proceedings of the 26th European Conference on Information Retrieval*, 57-71.

[2] Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*. **56**(1): 71-90.

[3] Brajnik, G., Mizzaro, S., Tasso, C., and Venuti, F. (2002). Strategic help for user interfaces for information retrieval. *Journal of the American Society for Information Science and Technology*. **53**(5): 343-358.

[4] Busha, C.H. and Harter, S.P., (1980). *Research methods in librarianship: Techniques and interpretation*. Library and information science series. New York: Academic Press.

[5] Campbell, I. and Van Rijsbergen, C.J. (1996). The ostensive model of developing information needs. *Proceedings of the 3rd International Conference on Conceptions of Library and Information Science*, 251-268.

[6] Harman, D., (1992). *Relevance feedback and other query modification techniques*. In *Information retrieval: Data structures and algorithms*. New York: Prentice-Hall.

[7] Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference. *SIGIR Forum*. **37**(2): 18-28.

[8] Koenemann, J. and Belkin, N.J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 205-212.

[9] Meddis, R., (1984). *Statistics using ranks: A unified approach*. Oxford: Basil Blackwell, 303-308.

[10] Morita, M. and Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 272-281.

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

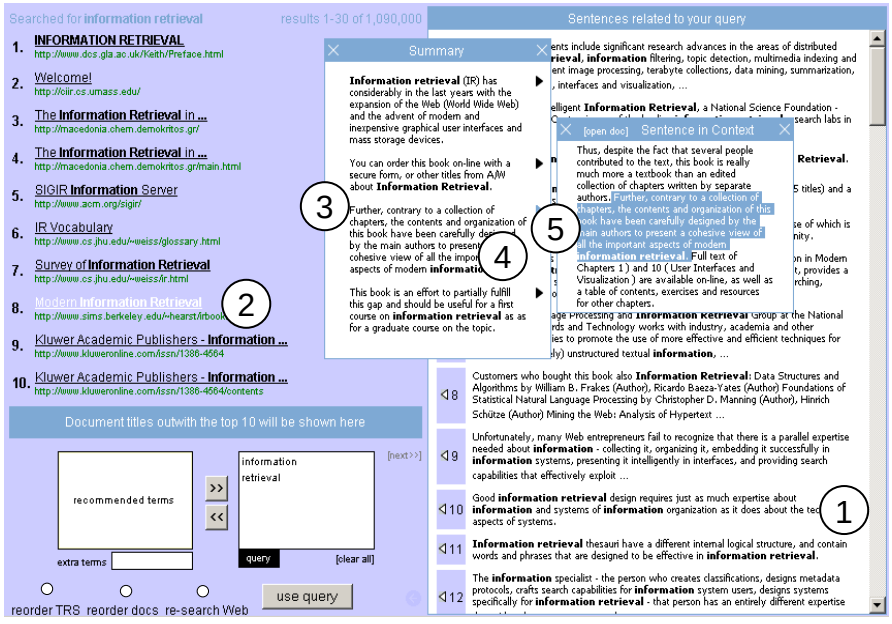
[12] Siegel, S. and Castellan, N.J. (1988). *Nonparametric statistics for the behavioural sciences*. 2nd ed. Singapore: McGraw-Hill.

[13] White, R.W. (2004). *Implicit feedback for interactive information retrieval*. Unpublished Doctoral Dissertation, University of Glasgow, Glasgow, United Kingdom.

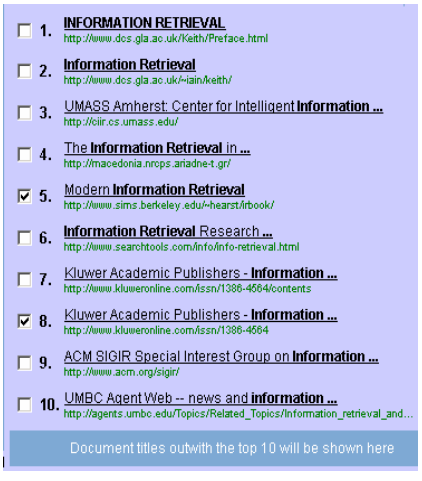
[14] White, R.W., Jose, J.M. and Ruthven, I. (2005). An implicit feedback approach for interactive information retrieval, *Information Processing and Management*, in press.

[15] White, R.W., Jose, J.M., Ruthven, I. and Van Rijsbergen, C.J. (2004). A simulated study of implicit feedback models. *Proceedings of the 26th European Conference on Information Retrieval*, 311-326.

[16] Zellweger, P.T., Regli, S.H., Mackinlay, J.D., and Chang, B.-W. (2000). The impact of fluid documents on reading and browsing: An observational study. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 249-256.



Appendix A. Interface to Implicit RF system.
1. Top-Ranking Sentence 2. Title 3. Summary 4. Summary Sentence 5. Sentence in Context



Appendix B. Checkboxes to mark relevant document titles in the Explicit RF system.