A Study of Interface Support Mechanisms for Interactive Information Retrieval

Ryen W. White [1]
Institute for Advanced Computer Studies
University of Maryland
College Park, MD USA 20742
ryen@umd.edu
Tel: +1 301 405 2748

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

Abstract

Advances in search technology have meant that search systems can now offer assistance to users beyond simply retrieving a set of documents. For example, search systems are now capable of inferring user interests by observing their interaction, offering suggestions about what terms could be used in a query or reorganizing search results to make exploration of retrieved material more effective. When providing new search functionality, system designers must decide how the new functionality should be offered to users. One major choice is between offering automatic features that require little human input but give little human control, or interactive features, which allow human control over how the feature is used but often give little guidance over how the feature should be best used. This article presents a study in which we empirically investigate the issue of control by presenting an experiment in which subjects were asked to interact with three experimental systems that vary the degree of control they had in creating queries, indicating which results are relevant and in making search decisions. We use our findings to discuss why and how the control users want over search decisions can vary depending on the nature of the decisions and the impact of those decisions on the user's search.

Keywords

Information retrieval, user interfaces, evaluation, implicit relevance feedback.

[1] Corresponding author.

1. Introduction

The widespread use of commercial search systems has highlighted the importance of user interaction in Information Retrieval (IR). Web search engines such as Google, Yahoo! and MSN Search have grown in popularity and process millions of queries daily. The users of these systems are responsible for all aspects of their interaction, from the selection of query terms to the assessment of the results obtained. This can be problematic as users typically receive no formal training in how to formulate queries, exhibit limited interaction with the results of their searches and do not examine results closely (Jansen, Spink & Saracevic, 2000).

The decisions that users must make can be divided into those related to three major search activities:

- *selecting query terms and operators*. Almost all search systems require a user to enter a set of terms to initiate a search. Some search systems also allow users to enter special operators such as "" to indicate phrases and +/– to indicate the inclusion/exclusion of terms. Boolean operators such as AND, OR and NOT can also be used to control search results. Search engines such as AltaVista have also been effective in suggesting new query terms to users (Anick, 2003).
- *making search decisions*. If a search does not retrieve all the relevant information required then a user must decide whether or not to continue with the search. The user has control over when new queries are issued, how many are issued and at what point to stop searching. Desktop search systems such as Yahoo! Desktop Search allow users to reorder retrieved search results based on criteria other than relevance to the query (e.g., modification date, file location). These information seeking decisions are controlled by users and fit within an overall search strategy.

- *indicating relevance*. If the user finds interesting or relevant pages the system can use the content of these pages to improve the search. Google's "similar pages" option is one example of this. Non-web search systems use *Relevance Feedback* (RF) (Salton & Buckley, 1990) which automatically modifies queries based on relevance information provided to the search system by the user. Browse-based interfaces can also use relevance information (what pages the user has selected to view) as the basis for deciding what new pages to offer. In relevance feedback the user must decide which documents to mark as relevant and when to ask the system to modify their query.

In all these areas the user must make decisions on some aspect (selection of search terms, continue searching, marking documents relevant, etc.) and leave some decisions to the system (how search terms are used to retrieve documents, how to modify the user's query, etc.). The result is that there is a shift in control from user to system and back again throughout the search. The decision on who has control over what aspects of the search is taken by the system designer who allows the user control over some decisions and the system control over others. However this balance does not necessarily reflect the way that users may wish to interact with search systems; they may want control over some aspects and less control over others.

This article presents an investigation of user control in the three aspects of searching outlined above: query creation, indicating relevance and search continuation. Our subjects perform searches on systems with different ways of doing each of these. The aim of this study is to establish how much control users *actually want* over each task. Through studies of this nature we can gain a better understanding of how search systems are used and how interface support in such systems should be offered.

This article is structured as follows: Section 2 motivates the work and we describe the experimental methodology in Section 3. In Section 4 we present the findings of our study, discuss their implications for the design of new search interfaces in Section 5 and conclude in Section 6.

2. Background and Motivation

Search interfaces are the means through which users interact with search systems and control aspects of their search. The importance of developing effective interfaces for these systems has already been widely acknowledged, (e.g., Hearst, 1995; Shneiderman, Byrd & Croft, 1998). However, in situations where users struggle to exercise their control effectively it may be desirable to delegate control over certain aspects of the search to the search system (Bates, 1990). It is vital therefore to determine when the system should take control and when the user should be involved (Beaulieu & Jones, 1998).

RF systems typically have components that allow users to indicate which information is relevant, create and modify queries and use these queries to update the results display. RF techniques have been shown to improve search effectiveness in non-interactive settings (Buckley, Salton & Allan, 1994) but the user interface challenge is to provide an easy and effective way to control their use in feedback systems; this is the focus of the experiment described in this article.

2.1 Indicating Relevance

In some search engines and RF systems users are responsible for explicitly indicating which documents are relevant. This technique allows users who may be unable to create effective queries to receive assistance by providing examples to the search system of what information is relevant. The information can then be used to retrieve similar documents or rank similar documents higher than dissimilar ones. RF is based on the principle by which users can describe

the property "relevance" to a search system by showing the system examples of search results that contain relevant information. RF gives users control over which information the search system regards as relevant but intrudes on their primary line of activity (i.e., finding useful information) and increases the cognitive burden (Beaulieu & Jones, 1998). That is, RF forces users to make two sets of decisions: decisions on finding relevant material and decisions on how to operate RF.

Giving users control over which documents the system uses is only one way in which relevance information can be provided. Systems that use *Implicit Relevance Feedback* (IRF) (Morita & Shinoda, 1994; Kelly & Teevan, 2003) make assumptions about the relevance of top-ranked search results or results with which users 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) (Kelly & Teevan, 2003) or using interaction with browse-based result interfaces (Campbell & Van Rijsbergen, 1996). These techniques have demonstrated that it is possible to elicit feedback from users in ways other than the traditional RF model. **In this study, we use implicit and explicit feedback techniques to investigate how much control users want over the process of indicating which search results are relevant.**

2.2 Creating Queries

Users are traditionally responsible for modifying their own queries during a search. However, this process of *query formulation* can be problematic if users have insufficient knowledge of the domain, search system or vocabulary used to index documents to create well-formed queries (Furnas, Landauer, Gomez & Dumais 1987; Salton & Buckley, 1990). When needs are ill-defined users may also face problems in transforming their need from one which they are consciously aware (i.e., their *conscious* need) to a search query for presentation to the system (i.e., their *compromised* need) (Taylor, 1968). RF systems attempt to solve both of these

problems by selecting alternate query terms from the information marked as relevant. In traditional RF systems the documents marked are used by the system to construct a new search query. The level of user involvement in this aspect of the feedback process can vary. That is, users can delegate all responsibility for creating new queries to the search system or retain full control over which terms to select (i.e., only use the RF to suggest alternative query terms). A number of studies have found that users exhibit a desire for RF and, in particular, term suggestion features (Hancock-Beaulieu & Walker, 1992; Koenemann & Belkin, 1996; Beaulieu, 1997). However, the evidence from these and related studies have indicated that the features of RF systems are not used in interactive searching (Beaulieu, 1997; Belkin et al., 2001; Ruthven, Tombros & Jose, 2001); there appears to be inconsistency between what subjects say they want and what they actually use when confronted with RF systems. **In this study, we use three techniques for selecting query terms to investigate the control users want over the query formulation process.**

2.3 Making Search Decisions

Once a new query is created it must be presented to the system and used in some way. Search systems typically use the new query to retrieve a new set of search results. Users are responsible for this aspect of the search and have control over when this decision is made. This is only one use of the query and it is also possible for search systems to choose when and how the modified query is used (White, Jose & Ruthven, 2005). Systems such as PLEXUS (Vickery & Brooks, 1987) and I^3R (Croft & Thompson, 1987) allow the selection of retrieval strategies by users or on behalf of users, but are dependent on the initial construction of a model to represent the user and their needs. In these systems users can select precisely what action happens and when it happens. **In this study, we use techniques that vary user control over how the query is used to investigate how much control users want over the selection of search decisions.**

The study described in this article uses systems that vary each of these three factors and asks users for their opinions of each variation to better understand what users of feedback systems actually want. We now provide more details on the study.

3. Study

The aim of the experiment on which this study is based was to investigate the effectiveness of three RF systems for interactive search. It is important to point out that the aim of this study is not to test the effectiveness of new interface techniques, but to establish user preferences that may shape the future development of search systems. We were motivated by trends in our experimental results to pursue the investigation we present here. In this section we present details of the experiment performed, beginning with the experimental systems used in this study.

3.1 Experimental Systems

We first describe interface features common to all systems, then the differences between systems.

3.1.1 Interface Features

INTERFACE

The interface used in all experimental systems (Figure 1) allowed users to interact with the retrieved information at a lower level than traditional result presentation methods. The interface uses many representations of the same Web pages (documents) to allow different views on the information contained within documents and has been shown to be more effective than traditional forms of Web search result presentation (White, Jose & Ruthven, 2003a).

[PUT FIGURE 1 HERE]

DOCUMENT REPRESENTATIONS

Documents are represented at the search interface by their full-text and a variety of smaller, query-relevant representations, created at retrieval time. The top-30 retrieved documents are downloaded and all sentences from each document are extracted[2]. Each sentence is assigned a score, using an algorithm similar to that in (White, Jose & Ruthven, 2003b). This algorithm gives preference to sentences that contain the user's query terms. These sentences are used to form many representations of each document. Interacting with a representation guides users to a different representation from the same document.

Document representations include the document title (2)[3] and the query-biased summary of the document (3); a list of *top-ranking sentences* (TRS) extracted from the top-30 documents retrieved, scored in relation to the query (1), a sentence in the document summary (4), and each summary sentence in the context it occurs in the document (i.e., with the preceding and following sentence) (5). Each summary sentence and top-ranking sentence is regarded as a representation of the document. These representations allow users to more deeply explore the retrieved information and can combine to form an interactive *relevance path* at the search interface. The default display contains the list of top-ranking sentences and the list of the first ten document titles. Interacting with a representation guides users to related representations from the same document. If they click the arrows next to the numbers adjacent to the top-ranking sentences the system highlights the title of the source document. If they hover over a document title for a short time the summary of that document appears in a small, moveable window in front of the other information. Clicking arrows next to sentences in that summary shows the sentences in the context they occur in the source document. The presentation of progressively more information

[2] This number of documents was chosen to ensure the system responded in a timely manner.
[3] Numbers correspond to Figure 1.

from documents to aid relevance assessments has been shown to be effective (Zellweger, Regli, Mackinlay & Chang, 2000; Paek, Dumais & Logan, 2004).

All experimental systems contain components to help users construct improved search queries. Once created, the query can be used in different *search decisions* to generate a new set of results (i.e., re-search Web) or restructure already retrieved information (i.e., reorder top-30 documents or list of top-ranking sentences).

In this study we use different versions of the system that vary how users indicate which document representations are relevant, modify their queries and make search decisions. In the remainder of this section we describe these three variations.

3.1.2 Manual System

This system allowed users to indicate directly which document representations were relevant until they were satisfied with the information marked. There are checkboxes next to all document representations (including sentences and summaries) and using these the user can mark representations as relevant; this is effectively a standard RF interface. Figure 2 shows an example of this at the interface. In the figure, checkboxes next to each title allow users to mark titles as relevant.

[PUT FIGURE 2 HERE]

The interface contains control options that allow the user to request support with query formulation, modify the query and choose retrieval strategies. The options, shown in Figure 3, appear in the bottom left-hand corner of the interface.

[PUT FIGURE 3 HERE]

When the user is satisfied with the representations marked relevant, they can click the "create query" button and a new query will be constructed from the marked representations. Suggestions for query modification are generated by analysing the documents or representations that are marked as relevant. The terms chosen to expand the query are the best terms chosen by the RF algorithm for selecting new query terms described in (White, 2004). These terms are appended to the original query on a new line and presented in the search box for the user to edit (Figure 3).

In the Manual system the user has control over the nature and timing of search decisions (i.e., when to reorder the sentences, reorder the documents or re-search the Web). To do this, the user selects the radio button that matches their desired action and click the "use query" button.

3.1.3 Assisted System

In this system there are no checkboxes for users to explicitly mark what representations are relevant. Instead, this system makes inferences about users' interests based on the information with which they interact. As described earlier, interacting with a representation indicates other representations from the same document that may be displayed at the interface. To users this is a way they can find out more information from a potentially interesting source. To the system each interaction of this nature is interpreted as an implicit indication of interest in the representation and each representation is treated as relevant for the purpose of creating a list of new query terms. A version of this technique is described and has been shown to be a good estimation of explicit user feedback in (White, Ruthven & Jose, 2002).

Every five relevance paths (i.e., when a user views one or more representations from five separate documents), the system chooses a new set of keywords and search decisions based on the

system's estimate in the level of change in the topic of the search since the last user-controlled query submission. The system chooses the top-ranked terms and presents these in a list of recommended terms. The control options for this experimental system are shown in Figure 4.

[PUT FIGURE 4 HERE]

The user can then control which terms are added or removed from the query. The arrowed buttons can be used to transfer terms between the recommended list and the query. There is an "extra terms" box where users can add additional terms to the query that are not in recommended list.

When the system has a recommendation it shows its recommended terms and highlights the radio button for the search decision it recommends. The user does not have to agree with the recommended search decision and can choose another option.

3.1.5 Automatic System

The Automatic system obtains its relevance information implicitly in the same way as the Assisted system. However, the system retains control of all other choices (i.e., the query terms chosen and search decisions made). Rather than recommending what the user should do, the Automatic system chooses terms and makes search decisions without direct user instruction, then notifies the user.

This system allows the user to edit their original query and retrieve a new set of documents. No provision is made for them to formulate a query for reordering sentences or documents; these actions were controlled by the system. The system chose alternative query terms automatically and makes a search decision on the user's behalf.

This system notified users that a new set of documents had been retrieved or the already retrieved information was restructured using notification messages similar to that shown in Figure 5 in the bottom left-hand corner of the interface.

[PUT FIGURE 5 HERE]

The messages give details of the query terms chosen by the system and the action the system has taken. The query terms shown in the notification window are the full query not the list of terms appended to the initial query. It is possible therefore for the new query to not contain the original query terms. This design decision allows the Automatic system to automatically adapt to large changes in the topic of the search without being tied to terms in the initial query.

3.2 Summary of Systems

In Table 1 we present a summary of the role of the user in indicating relevance, constructing queries and choosing how these queries are then used in each of the three experimental systems used in this study.

[PUT TABLE 1 HERE]

Subjects using the Manual system have control over relevance assessment. This system requires users to make binary relevance judgements (i.e., relevant/non-relevant). Although there exist ways of eliciting degrees of relevance from users at the interface (Ruthven, Lalmas & Van Rijsbergen, 2002) the need to make many assessments meant the binary approach was the least overwhelming and time consuming explicit RF alternative available to us. All experimental

systems allowed users to reverse the effects of re-searching or document/sentence reordering. In the next section we describe the experimental subjects who participated in this study.

3.3 Subjects

The experimental subjects were mainly undergraduate and postgraduate students in the Arts, Sciences and Social Sciences faculties at the University of Glasgow, United Kingdom. 48 subjects were recruited; half were male and half were female. Recruitment was targeted at two subject groups, each containing 24 subjects: *inexperienced* (infrequent computer users, inexperienced searchers) and *experienced* (frequent computer users/professional computer users, experienced searchers). Subjects completed an entry questionnaire that asked them about their search experience and computer use. They were then divided into the two groups depending on their search experience, how often they searched and the types of searches they performed.

The average age of the subjects was 22.83 years (maximum 51, minimum 18, $\sigma = 5.23$ years) and three quarters had a university diploma or a higher degree. 47.91% of the 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. All had some degree of experience with Web searching[4], and some with searching in other domains[5].

3.4 Tasks

Search tasks were designed to encourage realistic search behaviour by our subjects and were search scenarios that reflected real-life search situations. The tasks were phrased in the form of simulated work task situations (Borlund, 2000), i.e., short search scenarios that were designed to reflect real-life search situations and allow subjects to develop personal assessments of relevance.

[4] Inexperienced subjects conducted Web searches on average "Once or twice a week", Experienced subjects conducted Web searches on average "Many times a day".

[5] Examples include: the University of Glasgow library, the British Library, their personal computer with desktop search tools.

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 Bell and Ruthven (2004) described, 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 RF algorithms. Task complexity was varied according to the methodology described by (Bell & Ruthven, 2004), specifically by varying the number of potential information sources and types of information required to complete a task.

Subjects chose one high complexity, one moderate complexity and one low complexity task. They chose a task from a different topic each time and were not allowed to choose more than one task for a particular topic. This minimised learning effects. Giving subjects a choice of topics allowed them to select those that were most interesting. Borlund (2000) argues that interest is one of the key factors in engaging subjects in simulated work task situations. The three tasks devised for the "Petrol Prices" topic are shown as an example in the Appendix. 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. That is, highly complex tasks can encourage exploratory searching (e.g., browsing) and simple tasks focused directed searching (e.g., keyword search) (Kuhlthau, 1993)[6]. In the next section we describe the experimental procedure.

[6] For succinctness of exposition we do not use the differences in complexity in our analysis. A detailed description of the role of task complexity in this study can be found in (White, 2004).

3.5 Procedure

The experiment has a 2 × 3 factorial design; two levels of search experience and three experimental systems. Subjects switched systems after each task and used each system once. The order in which systems was used was randomised according to a Latin square experimental design. 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 five 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 the first task sheet and asked to choose one task from the six on that sheet. No guidelines were given to subjects when choosing a task although complexity was rotated by the experimenter so each subject attempted one high complexity task, one moderate complexity task and one low complexity task.

v. after selecting the task, the subject was asked to perform the search and was given 15 minutes to search. Subjects 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 task sheets were given to the subject, following steps iv. – vi. Since the topics were the same on all three task sheets the subject was not allowed to choose the same topic as attempted in a previous search even though subsequent choices would be from a different level of complexity.

viii. the subject completed a post-experiment questionnaire and participated in a post-experiment interview.

The findings of this study are now presented.

4. Findings

In this section we mainly focus on results concerning the interface differences between systems. This is a study of how much control users want over aspects of their search, not of new techniques to improve search effectiveness. As such, the findings we present focus mainly on subjective impressions of the interface support mechanisms of each system.

Due to the ordinal nature of much of the data non-parametric statistical testing was used and the level of significance was set to $p < .05$. The findings are presented across three aspects of the search: indicating relevance, creating queries and interactive search strategies. S_{Man}, S_{Assist} and S_{Auto} are used to denote the Manual, Assisted and Automatic systems respectively. We used 5-point Likert scales and semantic differentials, and open-ended questions to elicit subject opinion (Busha & Harter, 1980). System logging was also used to record subject interaction.

4.1 Indicating Relevance

The experimental systems differ in how users convey which items are relevant. The Manual system presents checkboxes next to each document representation and allows users to explicitly mark which representations are relevant. The ability to mark items as relevant gives users an increased responsibility for making decisions but more control over the input to the system and when system operations are carried out. Relevance indications on the Assisted and Automatic systems are implicit. That is, the systems make inferences about what information is relevant from their interaction. In this section we analyse subject perceptions of these methods.

4.1.1 Subject Perceptions

Subjects were asked about how relevance information was conveyed to each of the systems. That is, how they told the system which document representations (e.g., titles, summaries, top-ranking sentences) were relevant. They were asked to complete semantic differentials to elicit subject opinion about:

1. the *value* of the assessment method i.e., *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 i.e., *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 2 for inexperienced subjects, experienced subjects and all subjects regardless of search experience. The value corresponding to "All" represents the mean of all differentials for a particular attitude statement (e.g., all three differentials for statement 1). This gives some overall understanding of the subjects' feelings which can be useful as subjects may not answer individual differentials very precisely. Bold font

is used in this table and in all subsequent tables to denote the highest (or most positive) value for a particular combination of variables (e.g., "easy"/*inexperienced*, most positive is S_{Auto} (1.79)).

[PUT TABLE 2 HERE]

Friedman Rank Sum Tests were applied within each subject group[7]. The results of this analysis suggested the existence of significant differences in all semantic differentials and all subject groups[8] except the "comfortable"/*experienced* comparisons[9] (underlined in Table 2). Experienced subjects appear equally comfortable with the methods used to provide relevance information in all systems. Their search experience may allow them to move between interface technologies more easily.

Dunn's *post hoc* tests (multiple comparison using rank sums) were run on all differentials revealing significant differences for all pair-wise comparisons. These differences suggest that subjects found the implicit feedback methods "easy", "effective" and "useful" in their search[10]. In the Manual system subjects could decide which document representations were marked as relevant. Subjects felt more "in control" when given the additional responsibility for indicating which items were relevant but, for inexperienced subjects, not necessarily more comfortable. Users with less search experience may find it problematic to adapt to new techniques for controlling their search. All subjects found (explicit) relevance assessment in Manual system more difficult than (implicit) assessment in the Assisted and Automatic systems. However, the significance of the difference between the S_{Assist} and S_{Auto} systems suggests that factors other than

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

[8] all $\chi^2(2) \geq 10.60$, all $p \leq .005$

[9] $\chi^2(2) = 2.94$, $p = .23$

[10] all $Z \geq 2.26$, all $p \leq .012$

the value and the process of relevance indication affect subject preferences for different relevance assessment methods.

4.1.2 Search Precision

We also use analysis of interaction logs to investigate how subjects actually conveyed relevance in our experimental systems. To do this we measure the *precision* of the search; the proportion of all possible representations in the top-30 documents retrieved that were relevant. In the Manual system the search precision is in two forms: (i) the proportion of *all possible* representations that were marked relevant by the user, and (ii) the proportion of *all viewed* representations that were marked relevant by the user. In the Assisted and Automatic systems precision is based on the proportion of all possible representations that the user expresses an interest in (i.e., viewed). 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 to all three systems shows document representations from the top-30 documents there are 420 possible representations that subjects can assess. Table 3 shows precision values for each system. For the Manual system, the precision value is given in the format: precision from *all possible* representations, (precision from *all viewed* representations) (potential precision if implicit feedback had been used).

[PUT TABLE 3 HERE]

The average search precision values shown in Table 3 suggest large differences in the number of items marked relevant in the Manual system and inferred relevant. Subject criteria for marking a representation was generally very strict. The Manual precision values differ significantly from

those of the Assisted and Automatic systems for both subject groups and overall[11]. The precision values for the Assisted and Automatic are very similar and do not differ significantly between subject groups[12]. From these results we can see that experienced subjects check more items yet look at fewer. This may imply that they are being selective about the information they view, but apply different criteria than inexperienced subjects when assessing relevance.

The results of this analysis indicate differences in feedback users are willing to give and the amount that can be gathered implicitly. In the next section we present subject opinions of this process.

4.1.3 Subject Opinions

Subjects were asked to comment informally on each of the experimental systems they used. Subjects found the Manual system a hindrance in their search, that it presented them with too many choices and that it added an additional component to the search process that could become frustrating. Subjects found the need to mark representations in the Manual system annoying and reduced its usability. Three factors emerged as important when indicating which results were relevant: the method used to indicate, the value of the indication and the criteria used during the indication. The *method* describes how relevance information was elicited at the interface and the subjects typically forgot to provide these indications. The *value* describes the perceived benefit of indicating relevance and subjects generally felt the process was not worth their effort. Finally, the *criteria* that the subjects employed during the process were typically strict (i.e., results had to be completely relevant) and subjects rarely found results they regarded as relevant. These findings demonstrate the need for functional visibility in the RF process and the ways to address the high cognitive load imposed by explicit RF systems (Beaulieu and Jones, 1998). The Automatic and Assisted systems provided a mechanism through which relevance information

[11] Wilcoxon Signed-Rank Test, all $T(24) \geq 229$, all $p \leq .012$
[12] Mann-Whitney Test, $U(24) = 351$, $p = .097$

could be conveyed that was found to be straightforward and did not disrupt subjects' search patterns. From their comments subjects appeared willing to delegate responsibility for this activity to the search system.

4.2 Creating Queries

At any point in the search the experimental systems allowed new search queries to be created. When prompted, the Manual system presented the original query and the best non-query terms in a text box and allowed the user to add additional terms or remove terms to formulate the new query. The Assisted system presents a list of recommended terms and allows the user to add the best from this list into the query or remove terms from the query. The Automatic system generated a new non-editable query automatically. The Manual and Assisted systems gave users control over their query terms. In this section we present subject perceptions of this process and the levels of subject trust in the systems to form new queries.

4.2.1 Subject Perceptions

Subjects were asked to indicate on a Likert scale how comfortable they were with the method for constructing the new query. The average responses are shown in the third row of Table 4.

[PUT TABLE 4 HERE]

A Friedman Rank Sum Test was applied to the values in each group and the results indicated statistically significant differences in all groups[13]. Dunn's *post hoc* tests were applied to the data and revealed (in all groups) significant differences between the Assisted system and the other

[13] all $\chi^2(2) \geq 17.03$, all $p < .001$

systems[14]. The differences between the Manual and Automatic systems were not significant in any groups[15].

The subjects appear to prefer systems that recommended terms to them in a way that does not intrude on their search, giving them control over which terms could be added to their query. The adding of terms represented an additional burden, but did not lessen their perceptions of the technique.

4.2.2 Subject Trust

Trust is an important factor when relying on others. A relationship between user trust and willingness to use controlling mechanisms or accept automated assistance has already been established in the Ergonomics community (Lee & Moray, 1994). The same principles can be applied to Interactive IR; to delegate responsibility to a search system, users must be able to trust the system to act on their behalf. Subjects were asked whether they trusted the system to choose terms for them. They completed a Likert scale to indicate the extent they agreed with the statement *I would trust the system to choose terms for me*. The last row of Table 4 shows the average responses.

Friedman Rank Sum Tests were conducted for each differential within each subject group. The results suggested the existence of statistically significant pairs[16]. Dunn's *post hoc* tests revealed significant differences in all inexperienced comparisons and for the experienced and overall subject groups, the Assisted /Automatic[17] and Assisted/Manual[18]. Subjects appear to trust systems that give them control over query modification more than those without this facility. Subjects are

[14] all $Z \geq 3.12$, all $p < .001$
[15] *inexperienced*: $Z = 1.16$, $p = .123$; *experienced*: $Z = 1.08$, $p = .141$, *overall*: $Z = 1.08$, $p = .139$
[16] all $\chi^2(2) \geq 11.24$, all $p \leq .001$
[17] *experienced*: $Z = 2.03$, $p = .021$; *overall*: $Z = 2.00$, $p = .023$
[18] *experienced*: $Z = 2.05$, $p = .020$; *overall*: $Z = 1.90$, $p = .029$

more willing to delegate responsibility for the creation of queries to systems that allow them to verify the correctness of system decisions. In a related study, Koenemann and Belkin (1996) tested search systems with different levels of visibility and interactivity in creating queries. In our study the Automatic system allows users to see the query created by the system. The Manual and Assisted systems allow users to control and adjust the new query. In our study, as in (Koenemann & Belkin, 1996), subjects prefer systems that gave them control over the new queries. That is, they want help in selecting query terms but want ultimately to decide which terms are used. From the results presented in Table 4, we can see that users prefer interface mechanisms that give them control over query contents.

4.2.3 Source of Additional Query Terms

In all systems users could modify their query at any point in the search. This would involve them selecting additional query terms based on tacit knowledge, the search task, their general search experience and any additional information provided by the search system. After each search task subjects were asked to describe the origin of all additional query terms they entered during the search. These were not terms that the system suggested, but additional terms that users entered that may have originated in ideas the system terms gave them.

Subjects could select one from "list of terms suggested by the system", "retrieved set of documents and extracted information", "a combination of the first two" and "other". Subjects who chose "other", were asked to specify the reason. Table 5 shows the origins of new terms entered by the user. The values in the table are percentages and the sum of each column is 100%.

[PUT TABLE 5 HERE]

Most subjects appeared to choose additional terms based on the combination of the terms chosen by the system and the documents and extracted information. The abundance of information in these representations means it is more likely that new ideas will arise from there. What is encouraging is that the small number of terms selected by the system are not only useful to represent current information needs but to facilitate their development. Friedman Rank Sum Tests were conducted for each differential within each subject group. The results implied the existence of statistically significant differences in each group[19]. The high percentage of new ideas from "other" sources (the percentages shown in the last row of Table 5) came from a combination of the search task and the subject's tacit knowledge. The differences between the subject groups is significant for all differentials[20]. There is also evidence of interaction effects between the level of search experience and the experimental systems for the "combination of the above" and "other" differentials[21]. This suggests that the level of search experience affects where subjects get their terms and that this source varies depending on the experimental system.

The findings show that in systems that removed control over the generation of alternative query terms from subjects, they were more likely to use the terms proposed to initiate new ideas and search directions. The Manual system was dependent on subjects marking results as relevant. As a consequence, the terms suggested were from items the user already knew were relevant. Systems that remove subject control over creating queries may be most appropriate for encouraging new and potentially useful search directions. This can be helpful if the user is struggling with their search. Whilst users want to retain control over the additional terms used, if they are not experienced users, it may not be in their interests to do so.

[19] all $\chi^2(2) \geq 9.92$, all $p \leq .007$
[20] all $U(24) \geq 392$, all $p \leq .016$
[21] all $\chi^2(2) \geq 5.80$, all $p \leq .002$

The findings also show that the amount of interactivity in how additional terms were chosen influences where the terms were chosen from. When given less control over how alternative query terms were chosen, subjects were more likely use the system's terms or other sources such as the task, tacit knowledge or previous search experience. However, subjects did not use the documents or extracted information as inspiration for new terms. Subjects depend on the Automatic system to reorder documents and top-ranking sentences; subjects did not have any control over those activities in that system. We can conjecture that when subjects could not manipulate the space in which they searched, they were less likely to use that space to assist them in constructing new queries.

4.2.4 Subject Opinions

Subjects were asked informally about the activity of creating queries in each of the three experimental systems. Subjects preferred being able to select the terms used in the creation of their query. They did not like the Automatic system which did not let them refine their query for certain system operations. The selection of query terms is an activity for which users want support from the system in proposing additional keywords and suggested that this could be helpful where they may not be able to create good queries. However, subjects viewed the creation of new query as an important activity which they would rather have ultimate control over.

4.3 Making Search Decisions

Once a new query was created it could be used to retrieve a new set of documents, reorder top-ranked sentences or reorder documents. The Assisted and Automatic systems both contain a component that predicts when, and by how much, the topic of a search has changed. This component selects search decisions for execution or recommendation. In this section we analyse subject perceptions of these decisions.

4.3.1 Subject Perceptions

The experimental systems implemented retrieval strategies to gather a new set of documents or restructure the information already retrieved. The Automatic system acts on behalf of subjects, the Assisted system recommends a strategy and the Manual system is solely dependent on the subject to choose a strategy. In a similar way to the previous section, subjects were asked to indicate on a Likert scale how comfortable they were with the method used to make search decisions in the experimental systems. A summary of their responses is shown in Table 6.

[PUT TABLE 6 HERE]

A Friedman Rank Sum Test was applied to the values in each group and the results indicated the presence of effects in all groups[22]. Dunn's *post hoc* tests were applied to the data and revealed (in all groups) significant differences between all systems and all other systems (all $p \leq .001$). There were no significant differences between subject groups[23] and no significant interaction effects between search experience and systems[24]. Subjects preferred the Assisted and Manual systems since they had final control over how the new query was used. The Assisted system was preferred because it also made recommendations about possible uses of the query. The Automatic system was not liked because it removed this control and intruded on the search. The option to reverse all strategies did not compensate subjects for the additional burden of having to do so.

[22] all $\chi^2(2) \geq 14.26$, all $p < .001$
[23] Mann-Whitney Test, $U(24) = 350$, $p = .10$
[24] $\chi^2(2) = 1.94$, $p = .38$

4.3.2 Subject Trust

In the same way as the techniques used to create queries, trust is important in choosing how these new queries can be used. To effectively delegate responsibility users must be able to trust the systems to use the new query in the best possible way.

Subjects were asked about this aspect of the search. They completed a Likert scale to indicate the extent they agreed with the statement *I would trust the system to choose an action for me*. The average responses are shown in the final row of Table 6. Since the attitude statement concerned trust in *system decisions* it was not completed by subjects when they used the Manual system.

Wilcoxon Signed-Rank Tests were applied within each subject group to compare systems and all subjects and systems compared to the mid-value of the Likert scale (i.e., three). The results showed no significant within-group differences[25] and significant differences from the mid-value[26]. Subjects reacted positively to the search decisions made by the system. Inexperienced subjects preferred systems where they had control over the search decisions made. That is, they trusted systems that gave ultimate control over how the new query was used.

Another indicator of how much trust subjects had in the decisions made by the system is the proportion of decisions that subjects chose to reverse. This can be an indicator of dissatisfaction with the system and as an indicator of the extent to which subjects trusted the system to make the right choice on their behalf (i.e., the more decisions they reverse, the less they trusted the system to make the correct decisions for them). In Table 7 we present a summary of the proportion of search decisions made by the system that were accepted and those reversed by experimental subjects in each subject group and across all subjects, for each type of search decision.

[25] all $T(24) \leq 160$, all $p \geq .390$
[26] $T(24) = 229$, $p = .012$

[PUT TABLE 7 HERE]

The differences between the systems for all decisions within the subject groups is not significant[27] but it is between subject groups[28]. Experienced subjects tended to accept a lower number of search decisions than inexperienced subjects. These subjects may be more reticent about the search systems making decisions of this nature on their behalf and feel able to make such decisions on their own. In contrast, there were no significant differences in the proportion of decisions reversed between the three types of decision[29]. Subjects appeared equally satisfied and equally trustful of all types of search decision made by the systems.

4.3.3 Subject Opinions

Subjects were asked to comment informally about the search decisions. The Automatic system removed all user responsibility for making new decisions, the Assisted system recommended search decisions and the Manual system relied on users to make these decisions. In a similar way to how they felt for query creation subjects wished to retain control over the strategies employed, but responded well to recommendations made by the system. Where the retrieved information was restructured (i.e., reordering) rather than recreated (i.e., re-searching), subjects were more willing to delegate control to the search system. That is, the amount of control subjects wished to retain was dependent on the predicted impact of the search decisions.

We now discuss the results and their implications for search interface design.

5. Discussion and Implications

In this study we investigated interface support mechanisms for interactive information retrieval. The study focused on how much control users wished to retain over aspects of their search. A

[27] Wilcoxon Signed-Rank Test, all $T(24) \leq 156$, all $p \geq .431$
[28] Mann-Whitney Test, all $U(24) = 399$, all $p \leq .011$
[29] Friedman Rank-Sum Test, all $\chi^2(2) \leq 2.94$, $p \geq .23$

deeper understanding of what users want to control and what they are happy to delegate can assist in the development of more effective systems for interactive search.

Bates (1990) presented a framework for thinking about search system design that related *system involvement* in the search process and the *search activities* that systems directly support. System involvement ranges from Level 0 (i.e., no involvement) to Level 4b (i.e., complete system involvement with no user notification). Search activities include *moves* (identifiable thoughts or actions that are part of information seeking), *tactics* (one or more moves made to further a search), *stratagems* (large/complex sets of moves/tactics) and overall *strategy* that determines the direction of the search. The systems used in this study are involved in the search to different degrees; the Manual system suggests search activities when asked (i.e., Level 3a involvement in Bates' framework), the Assisted system offers search activities always (i.e., Level 3b involvement) and the Automatic system acts automatically and notifies user that it has done so (i.e., Level 4a involvement). All systems provide support for moves by allowing users to view documents and document representations and for tactics by providing assistance with query formulation and relevance indication. The Assisted and Automatic systems also provide assistance with stratagems by recommending or executing ways in which the query could be used such as reordering the top-ranked results or re-searching the document collection.

In this study we have used three systems that vary the level of support for tactics and stratagems. Meadows (1979) showed that it was problematic for systems to suggest *moves* as they may not be in line with the overall goals of the search. However, in our study the success of the systems using implicit relevance feedback do show that by tracking moves (e.g., document or document representation selections) during a search it is possible to build an approximation of user intentions, that can be used to recommend tactics and stratagems.

In a related study, Beaulieu and Jones (1998) investigated three factors that affect interaction with IR systems: functional visibility, cognitive load and balance of control between the user and system, relating them to a previous set of experiments. The functional visibility – allowing the user more information on how the system works – is important at two levels. Not only must the user be aware of what options are available at any stage but they must also be aware of the effect of these options. The study by Beaulieu and Jones demonstrated that interfaces such as the Manual system, that separate query modification and relevance assessment, can be more cognitively demanding for users. In this experiment subjects appeared willing to delegate responsibility for relevance assessment to the search system. However, they wished to retain control over query reformulation and retrieval strategy selection, activities they perceived as being important for the success of their search. That is, subjects were willing to delegate control over the provision of relevance information as long as they could control how this information was used.

A deeper understanding of what users want to control and what they are happy to delegate can assist in the development of more effective systems for interactive search. Techniques to indicate which items are relevant, form new queries and use these queries were all evaluated. In this section we discuss the findings of our evaluation for each of these techniques.

5.1 Relevance Indications

Subjects wanted the search system to infer relevance. In all cases, systems that gathered relevance information unobtrusively from subject interaction were preferred to systems that required explicit subject involvement. Whilst the Manual system gave subjects an opportunity to directly indicate which items were relevant the additional responsibility dissuaded subjects from doing so. They felt that the implicit techniques were a reasonable approximation for their indications and were willing to delegate responsibility for this activity to the search system.

Subjects felt that implicit relevance feedback was easier and more useable and that it was comparable in terms of search success.

The Manual system differed from the other systems in how relevance information was conveyed; the subject was required to explicitly mark representations as being useful in their search. This was an onerous task that was not liked by subjects. In the experiment one subject commented "[checking boxes] added a new dimension to search that could become frustrating". This summarises the general opinion of experimental subjects; that the need to mark boxes was removed from the search for information and required a transition between two search activities: locating useful information and marking that information if relevant. Subjects preferred systems that used implicit relevance assessments since they did not require them to mark items as relevant, they had difficulty marking items as relevant, they forgot to mark items and the marking of the items intruded in their searching. Implicit relevance assessments may not be as accurate as their explicit counterpart in determining which items are *definitely* relevant but they are able to build a larger body of evidence for those that are *potentially* relevant. The Manual system forced subjects to make binary assessments of what items were relevant; this may not always be appropriate as the relevance of a search result may be uncertain or partial (Spink, Griesdorf & Bateman, 1998; Maglaughlin & Sonnenwald, 2002).

Experimental subjects tended to only mark items that were definitely relevant, meaning they did not provide the system with much evidence with which to make query modification decisions (i.e., only around 2% of all representations were marked). Techniques such as those employed by Aalbersberg (1992), Allan (1996) and Iwayama (2000) can be used to modify queries in situations where only a small amount of relevance information is available. 15 of the 48 experimental subjects suggested that the process of relevance feedback could also be improved if they could provide indications of what interface items or terms definitely were not relevant for

their search. After they had given this negative relevance feedback they would not want to see items of this nature, or these terms, again during their search.

In this experiment "precision" was taken as a measure of search effectiveness and based on how much of the retrieved document set the subjects classed as relevant. To compute this measure, the Manual system used the proportion of potential representations[30] that were actually marked and the implicit feedback systems used the proportion of all representations that were classified as being relevant. The results suggested a large difference between how much information the implicit systems regarded as relevant and what the subject actually marked as being relevant. The relevance and usefulness of the terms generated from the implicit feedback systems was higher than that of the Manual systems, suggesting that more evidence, albeit less reliable than that provided by the user allowed better quality terms to be chosen by the implicit feedback framework. It also suggests that criteria subjects employed when assessing relevance was too strict and that better queries could have arisen from the selection of more representations that were perhaps not totally relevant.

5.2 Query Creation

Subjects preferred to retain control over query creation. The systems that allowed subjects to monitor and change the query were preferred over the Automatic system, which did not. They were willing to delegate the task of recommending potential keywords but not the task of adding these terms. Subjects preferred control over the terms chosen by the system, even if this meant more work for them in moving terms of interest from the recommended term list to the query. This effort was seen to be both *unnecessary* (subjects were not forced to do it) and *worthwhile* (subjects perceived a benefit from it). The implicit nature of the evidence captured may make the

[30] All document representations in the top 30 documents that could be marked *and* all document representations that were viewed.

search decisions of systems that use it unreliable and subjects may rather retain control to be sure of their correctness. Subjects engendered more trust in systems where they could verify the correctness of the terms chosen prior to their submission.

Subjects liked having terms suggested to them, but in a way that did not require them to delete irrelevant terms (as in the Manual system), only select relevant ones; subjects did not want to have to act to correct erroneous system decisions. Subjects were more willing to delegate responsibility for the creation of queries to systems that allow them to verify the correctness of system decisions. In a related study, Koenemann and Belkin (1996) tested search systems with different levels of visibility and interactivity in creating queries. In this study the Automatic system only allowed subjects to see the query created by the system; the Manual and Assisted systems allow subjects to view *and* adjust the new query. Here, as in Koenemann and Belkin, subjects preferred systems that gave them control over the new queries; they want help in selecting query terms but want ultimately to decide which terms are used. In this study we reinforce and extend Koenemann and Belkin's research to show that their findings are true across different types of searches.

The Manual system chose terms for subjects based on the items they had marked as relevant. These items reflected their current information needs and the terms suggested by the system appeared to reflect these needs also. Subjects chose terms from those recommended in the Assisted system because: (i) they represented new ideas, (ii) they meant the same as the query terms, and (iii) they were related to the query terms. The study by Koenemann and Belkin found that subjects tended to choose semantically related feedback terms. In this study we found that subjects use the query terms to give them ideas for what terms are appropriate or were related to the original terms in some way. For example, a search for "worldwide petrol prices" could mean

that the terms “pipe”, “iraq” and “dollar” are good feedback terms, but their semantic relationship to the original query is not immediately apparent.

All experimental systems tried to increase the length of subjects’ query statements by expanding the original search query. Belkin et al. (2003) have demonstrated that experimental subjects can be more satisfied with search results if they submit longer queries to the search system. The use of a feedback system to choose terms on a user’s behalf is only one way to create longer queries. It is preferable to encourage users to better define their information needs themselves. However, in circumstances where they may be unfamiliar with the topic of the search, they may be unable to produce longer queries (Kelly & Cool, 2002).

5.3 Retrieval Strategy Selection

Subjects preferred to retain control over search decisions. Systems that gave the subjects control over search decisions were preferred to those that did not. The Assisted system suggested decisions that subjects may execute. Subjects liked receiving this support but in a similar way to the creation of query statements wished to verify the correctness of any decisions before they were taken.

The Assisted and Automatic systems dynamically update their internal representation of information need change and adopt the search decision that reflects the level of change in the information need of the user, as estimated by the search system. Different search decisions had different levels of impact on a search. Reordering decisions restructured the already retrieved information at the interface, whereas re-searching decisions generated a new set of documents. The decisions increased in severity, from reordering Top-Ranking Sentences, to reordering documents, to re-searching the Web. Subjects appeared more willing to retain control over the number of re-search operations (63.22% of all accepted re-search decisions were initiated by the

user), but were willing to experiment with reordering (28.47% of all accepted reorder decisions were initiated by the user). This suggests an association between the severity of the decision and subjects' willingness to retain control over them. That is, for less severe strategies subjects were more willing to delegate responsibility to the system.

Different search decisions had different levels of impact on a search. Reordering decisions restructured the already retrieved information at the interface, whereas re-searching decisions generated a new set of documents. The decisions increased in severity, from reordering top-ranking sentences, to reordering documents, to re-searching the Web. Subjects appeared more willing to retain control over the number of re-search operations, but were willing to experiment with reordering. This suggests an association between the severity of decisions and subjects' willingness to retain control over them. That is, for less severe search decisions, subjects were more willing to let the system make the decision. As the search activity shifts from moves to tactics to stratagems (Bates, 1990), there is an increase in users' willingness to control them. Of these three activities, stratagems have the most influence on the success of the search; poor stratagem selection is potentially more serious than a mistaken move or tactic. Through allowing users to reverse the affects of bad stratagem selection our systems can handle incorrect decisions, although only if users are knowledgeable enough to notice that a bad decision has been made.

At the end of the experiment subjects were asked to rank the three systems they used in their order of preference. No instructions were given on what factors to use when making their decision, but they were asked to explain their ordering. Table 8 summarises the average rank assigned to each system.

[PUT TABLE 8 HERE]

The Assisted system received the highest ranking overall and for both subject groups, followed by the Manual system and the Automatic system, with significant differences between systems for both subject groups and overall across both subject groups[31]. The Assisted system received mainly positive comments and the Manual and Automatic systems mainly negative. The Manual system offers too many options, increased the burden on the subject and interfered with the process of finding information. Subjects generally felt that the Manual and Automatic system had good qualities: for the Manual system is it the control over which results are marked relevant, for the Automatic system it is the simplicity and control of the search. However, these qualities were insufficient to make subjects prefer these systems to the Assisted system. Subjects also generally felt more satisfied by relevance and usefulness of the terms suggested by the Assisted system and perceived the search decisions it made more positively than the other experimental systems we tested (White, 2004). The highest and the lowest ranked systems used IRF. Not only does this suggest that searcher satisfaction is not hampered by the use of IRF but that factors other than the RF method (e.g., query formulation and search decision selection) also affect searcher satisfaction.

Overall, the findings suggest that users want to retain control over the strategic aspects of their interaction i.e., over the aspects that will directly influence the quality of the results offered or future directions of their search. They view the provision of relevance indications as an operational activity required to receive assistance. There is a disparity between how important users regard the provision of relevance information and its importance to the search system. Although relevance feedback can be a useful tool to improve search effectiveness, it is underutilised because of the interface techniques it uses to gather relevance information. To cater for this, search systems must incorporate new techniques for gathering relevance indications. Implicit feedback methods similar to those described briefly in this article may be useful to

[31] Kruskal-Wallis Tests, all $\chi^2(2) \geq 4.61$, $p \leq .01$

address this problem. Further research is required in the development of search tools that incorporate implicit methods for capturing relevance information.

5.4 Limitations

It is worth mentioning two limitations of our study. First, the nature of the interfaces used for the experiment was specific for our purposes and was necessary to gather reliable implicit feedback to choose new query terms. However, further work is needed to test whether our findings can be applied to other types of search interface. Second, the process of making relevance assessments in the Manual system was complex and burdensome (subjects had to make *many* relevance assessments) and may have affected subjects' impressions of the usability of that system. However we do believe that the findings obtained add to the understanding of how interactive IR systems might usefully be designed.

6. Conclusions

In this article we have presented an investigation of user control in three aspects of the search typically supported by RF systems. We conducted a user study in which we tested different techniques for indicating relevance, creating queries and using these queries in different ways. Three experimental systems were developed that varied levels of control over each of these search aspects. Where appropriate we related our investigation to the work of Bates (1990), who addresses this issue of user/system control with respect to the level of system involvement and the search activity of the user or the system as part of the larger information search process.

We used the three experimental systems to investigate which activities users wished to retain control over, and how much control they actually wanted. Although we should be cautious about generalising our findings too much, the results of our study appear to show that users are happy to hand over full responsibility for indicating which search results are relevant, but only want to

receive assistance in the formulation of query statements and making search decisions. Users still wish to retain control over search activities they regard as important to the effectiveness of their search. Rather than trying to force users to provide feedback directly (as many RF systems do), IRF techniques can remove the burden of explicitly providing relevance information, allowing users to focus on those activities they regard as important.

7. References

Aalbersberg, I. J. (1992). Incremental relevance feedback. In Fox, E., Belkin, N., Ingwersen, P. and Pejtersen, A.M. (eds.) *Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 11-22). New York: ACM.

Allan, J. (1996). Incremental relevance feedback for information filtering. In Frei, H.-P., Harman, D., Schaübie, P. and Wilkinson, R. (eds.) *Proceedings of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 270-278). New York: ACM.

Anick, P. (2003). Using terminological feedback for web search refinement: A log based study. In Callan, J., Hawking, D. and Smeaton, A. (eds.) *Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 88-95). New York: ACM.

Bates, M. J. (1990). Where should the person stop and the information search interface start? *Information Processing and Management*, 25 (5), 575-591.

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

Beaulieu, M. and Jones, S. (1998). Interactive searching and interface issues in the Okapi best match retrieval system. *Interacting with Computers*, 10 (3), 237-248.

Belkin, N. J., Cool, C., Kelly, D., Lin, S.-J., Park, S.-Y., Perez-Carballo, J., et al. (2001). Iterative exploration, design and evaluation for query reformulation in interactive information retrieval. *Information Processing and Management*, 37, 403-434.

Belkin, N. J., Cool, C., Kelly, D., Kim, G., Kim, J. - Y., Lee, H. J., et al. (2003). Query length in interactive information retrieval. In Callan, J., Hawking, D. and Smeaton, A. (eds.) *Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 205-212). New York: ACM.

Bell, D. J. and Ruthven, I. (2004). Searchers' assessments of task complexity for web searching. In Macdonald, S. and Tait, J. (eds.) *Proceedings of the 26th BCS-IRSG European Conference on Information Retrieval*. (pp. 57-71). Berlin: Springer-Verlag

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

Buckley, C., Salton, G. and Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. In: Croft, W.B. and Van Rijsbergen, C.J. (eds.) *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 292-300). New York: ACM.

Busha, C. H. and Harter, S. P. (1980) *Research methods in librarianship: Techniques and interpretation*. New York: Academic Press Inc.

Campbell, I. and Van Rijsbergen, C. J. (1996). The ostensive model of developing information needs. In Aparac, T., Saracevic, T., Ingwersen, P. and Vakkari, P. (eds.) *Proceedings of the 3rd International Conference on Conceptions of Library and Information Science*. (pp. 251-268).

Croft, W. B. and Thompson, R. H. (1987). I^3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, 38 (6), 389-404.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30 (11), 964-971.

Hancock-Beaulieu, M. and Walker, S. (1992). An evaluation of automatic query expansion in an online library catalog. *Journal of Documentation*, 48, 406-421.

Hearst, M. (1995). TileBars: Visualization of term distribution information in full text information access. In Katz, R.B., Mack, R., Marks, L., Rosson, M.B. and Nielsen, J. (eds.) *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. (pp. 59-66). New York: ACM.

Iwayama, M. (2000). Relevance feedback with a small number of relevance judgements: incremental relevance feedback vs. document clustering. In Yannakoudakis, E., Belkin, N.J., Leong, M.K. and Ingwersen, P. (eds.) *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 10-16). New York: ACM.

Jansen, B. J., Spink, A. and Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36 (2), 207-227.

Kelly, D. and Cool, C. (2002). The effects of topic familiarity on information search behavior. In Hersh, W. and Marchionini, G. (eds.) *Proceedings of the 2nd ACM/IEEE Joint Conference on Digital Libraries*. (pp. 74-75). New York: ACM.

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

Koenemann, J. and Belkin, N. J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In Nardi, B., Van der Veer, G.C. and Tauber, M.J. (eds.) *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. (pp. 205-212). New York: ACM.

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

Lee, J. and Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153-184.

Maglaughlin, K. L. and Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgements. *Journal of the American Society for Information Science and Technology*, 53 (5), 327-342.

Morita, M. and Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In: Croft, W.B. and Van Rijsbergen, C.J. (eds.) *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 272-281). New York: ACM.

Paek, T., Dumais, S. T. and Logan, R. (2004). WaveLens: A new view onto internet search results. In: Dykstra-Erickson, E. and Tscheligi, M. (eds.) *Proceedings on the ACM SIGCHI Conference on Human Factors in Computing Systems*. (pp. 727-734). New York: ACM.

Ruthven, I., Tombros, A. and Jose, J. M. (2001). A study on the use of summaries and summary-based query expansion for a question-answering task. *Proceedings of the 23rd BCS-IRSG European Colloquium on Information Retrieval Research*. (pp. 1-14).

Ruthven, I., Lalmas, M. and Van Rijsbergen, C. J. (2002). Ranking expansion terms using partial and extensive relevance. In: Fidel, R., Bruce, H. , Ingwersen, P. and Vakkari, P. (eds.) *Proceedings of the 4th International Conference on Conceptions of Library and Information Science*. (pp. 199-219). Greenwood Village, CO: Libraries Unlimited.

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

Shneiderman, B., Byrd, D. and Croft, W. B. (1998). Sorting out search: A user-interface framework for test searches. *Communications of the ACM*, 41 (4), 205-212.

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

Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29, 178-194.

Vickery, A. and Brooks, H. M. (1987). PLEXUS: The expert system for referral. *Information Processing and Management*, 23 (2), 99-117.

White, R. W. (2004). *Implicit feedback for interactive information retrieval*. Unpublished doctoral dissertation, University of Glasgow, Glasgow.

White, R. W., Jose, J. M. and Ruthven, I. (2003a). In: Rauterberg, G.W.A., Menozzi, M. & Wesson, J. (eds.) A granular approach to web search result presentation. *Proceedings of the 9th IFIP TC13 Conference on Human Computer Interaction*. (pp. 213-220). IOS Press: Amsterdam.

White, R. W., Jose, J. M. and Ruthven, I. (2003b). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing and Management*, 39 (5), 707-733.

White, R. W., Ruthven, I. and Jose, J. M. (2002). The use of implicit evidence for relevance feedback in web retrieval. In: Crestani, F., Girolami, M. & Van Rijsbergen, C.J. (eds.) *Proceedings of 24th BCS-IRSG European Colloquium on Information Retrieval Research*. (pp. 93-109). Berlin: Springer-Verlag.

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. In: Szwillus, G. and Turner, T. (eds.) *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. (pp. 249-256). New York: ACM.

Appendix

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.

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.

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.

Figures

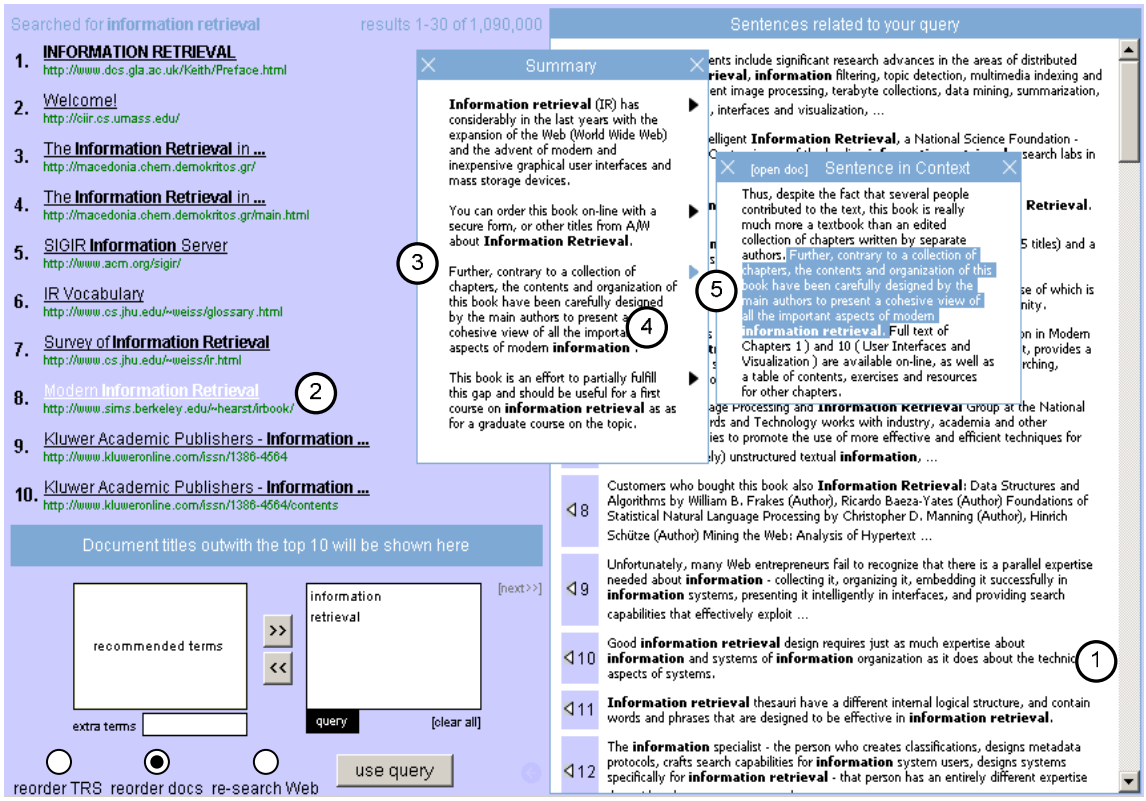

Figure 1. Search Interface (Assisted system)

Figure 2. Indicating relevance in Manual system.

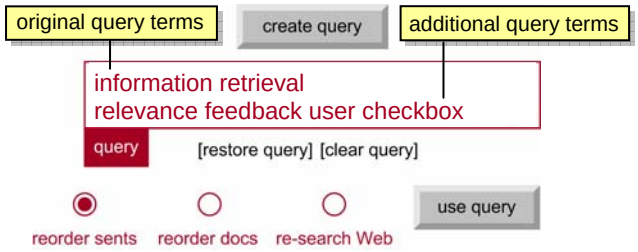

Figure 3. Control Options in Manual system.

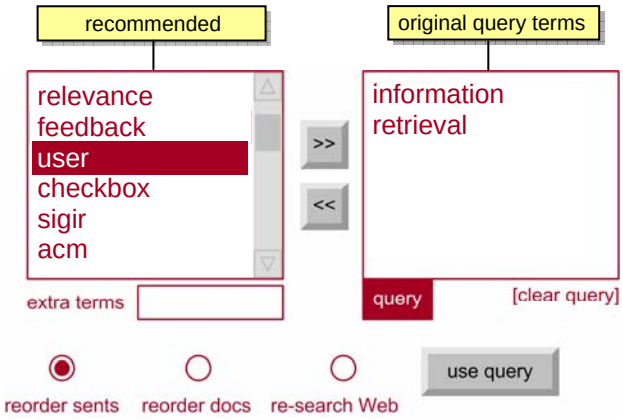

Figure 4. Control Options in Assisted system.

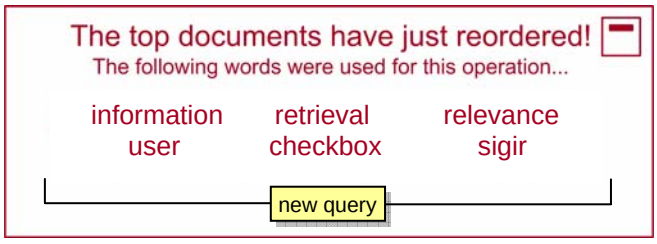

Figure 5. Automatic system notifications.

Tables

Table 1.User roles in experimental systems.

System	Manual	Assisted	Automatic
Relevance Indication	Control	Delegate	Delegate
Query Construction	Control	Delegate/Control	Delegate
Query Execution	Control	Delegate/Control	Delegate

Table 2. Subject perceptions of relevance indication (lower = better).

Differential	Inexperienced			Experienced			Overall		
	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}
Easy	2.46	1.88	**1.79**	2.46	2.00	**1.96**	2.46	1.94	**1.88**
Effective	2.75	**1.96**	2.67	2.63	**2.18**	2.67	2.69	**2.07**	2.67
Useful	2.50	**2.13**	2.42	2.46	**2.14**	2.40	2.48	**2.12**	2.41
All (1)	2.57	**1.99**	2.29	2.52	**2.11**	2.34	2.55	**2.05**	2.32
Comfortable	2.46	**1.88**	2.21	**2.14**	2.21	2.26	2.30	**2.05**	2.23
In control	**1.96**	2.25	3.21	**1.98**	2.13	3.14	**1.97**	2.19	3.13
All (2)	**2.21**	2.06	2.71	**2.06**	2.17	2.70	2.13	**2.12**	2.68

Table 3. Search precision (values are percentages).

Inexperienced			Experienced			Overall		
S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}
1.25 (5.96) (20.96)	21.65	21.36	2.76 (16.19) (17.05)	17.17	16.52	2.01 (10.57) (19.01)	19.41	18.94

Table 4. Subjective impressions of query creation methods (lower = better).

Differential	Inexperienced			Experienced			Overall		
	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}
Comfortable	2.79	**2.13**	2.96	2.63	**1.96**	2.88	2.71	**2.04**	2.92
Trust	2.19	**2.03**	2.48	2.19	**1.65**	2.19	2.19	**1.84**	2.34

Table 5. Origin of additional terms entered by the subject (values are percentages).

Source	Inexperienced			Experienced			Overall		
	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}
System terms	8.4	**20.8**	16.7	**29.2**	20.9	29.1	18.7	20.8	**22.9**
Documents and Extracted Information	20.8	**25.0**	16.7	29.2	**33.3**	16.7	25.0	**29.2**	16.7
Combination of the above	**50.0**	45.8	45.8	12.5	**33.3**	12.5	31.3	**39.6**	29.2
Other	**20.8**	8.4	**20.8**	29.1	12.5	**41.7**	25.0	10.4	**31.2**

Table 6. Subjective impressions of search decisions (lower = better).

Differential	Inexperienced			Experienced			Overall		
	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}
Comfortable	2.23	**2.04**	2.92	2.21	**1.94**	2.63	2.22	**1.99**	2.78
Trust	–	**2.67**	2.92	–	**2.67**	**2.67**	–	**2.67**	2.79

Table 7. Proportion of search decisions reversed (values are percentages).

Decision	Inexperienced		Experienced		Overall	
	S_{Assist}	S_{Auto}	S_{Assist}	S_{Auto}	S_{Assist}	S_{Auto}
Reorder sentences	**28.40**	**24.71**	35.21	30.95	**31.81**	27.83
Reorder documents	27.29	24.34	35.30	**31.48**	31.30	**27.91**
Re-search Web	27.03	24.16	**35.48**	30.27	31.26	27.22
All	**27.57**	24.40	**35.33**	30.90	**31.45**	27.65

Table 8. Rank order of systems (range 1-3, lower = better).

Inexperienced			Experienced			Overall		
S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}	S_{Man}	S_{Assist}	S_{Auto}
2.00	**1.45**	2.46	2.25	**1.29**	2.46	2.13	**1.42**	2.46