Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management xxx (2004) xxx–xxx
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

An implicit feedback approach for interactive information retrieval

Ryen W. White [a,*], Joemon M. Jose [a], Ian Ruthven [b]

[a] Department of Computing Science, University of Glasgow Glasgow, Scotland G12 8RZ, United Kingdom
[b] Department of Computer and Information Sciences, University of Strathclyde Glasgow, Scotland G1 1XH, United Kingdom

Received 20 February 2004; accepted 11 August 2004

Abstract

Searchers can face problems finding the information they seek. One reason for this is that they may have difficulty devising queries to express their information needs. In this article, we describe an approach that uses unobtrusive monitoring of interaction to proactively support searchers. The approach chooses terms to better represent information needs by monitoring searcher interaction with different representations of top-ranked documents. Information needs are dynamic and can change as a searcher views information. The approach we propose gathers evidence on potential changes in these needs and uses this evidence to choose new retrieval strategies. We present an evaluation of how well our technique estimates information needs, how well it estimates changes in these needs and the appropriateness of the interface support it offers. The results are presented and the avenues for future research identified.

Keywords: Implicit feedback; Interactive information retrieval; User evaluation

1. Introduction

A searcher approaches an information retrieval (IR) system with a need for information derived from an 'anomalous state of knowledge' (Belkin, Oddy, & Brooks, 1982). The subsequent transformation of this need into a search expression, or query, is known as *query formulation*. However, queries are only an approximate, or 'compromised' information need (Taylor, 1968), and may fall short of the description necessary to retrieve relevant documents. This problem is magnified when the information need is vague

* Corresponding author.
E-mail addresses: ryen@dcs.gla.ac.uk (R.W. White), jj@dcs.gla.ac.uk (J.M. Jose), ir@cis.strath.ac.uk (I. Ruthven).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2004.08.010

(Spink, Griesdorf, & Bateman, 1998) or searchers are unfamiliar with the collection makeup and retrieval environment (Furnas, Landauer, Gomez, & Dumais, 1987; Salton & Buckley, 1990). Consequently, search systems need to offer robust, reliable methods for query modification.

Relevance feedback (RF) (c.f. Salton & Buckley, 1990) is the main post-query method for automatically improving a system's representation of a searcher's information need through an iterative process of feedback. However, the technique assumes the underlying need is the same across all iterations (Bates, 1989) and relies on explicit relevance assessments provided by the searcher. These indications of which documents contain relevant information are used to create a revised query more similar to those documents marked relevant. The need to explicitly mark relevant documents means searchers may be unwilling or unable to directly provide relevance information (Beaulieu & Jones, 1998). Searchers may find the direct provision of relevance indications burdensome, cognitively demanding or they may be unable to identify what information is relevant.

Implicit RF, in which an IR system unobtrusively monitors search behaviour, removes the need for the searcher to explicitly indicate which documents are relevant (Morita & Shinoda, 1994). The technique uses implicit relevance indications, gathered unobtrusively from searcher interaction, to modify the initial query. Whilst not being as accurate as explicit feedback (since it removes searcher control), we have shown in previous work that implicit feedback can be an effective substitute for explicit feedback in interactive information seeking environments (White, Ruthven, & Jose, 2002b).

Traditionally, 'surrogate' measures such as document reading time, scrolling and interaction have been used to provide implicit evidence of searcher interests (Claypool, Le, Waseda, & Brown, 2001). However, such measures are context-dependent (Kelly, 2004), vary greatly between searchers and are hence difficult to correlate with relevance across searchers and searches. The relevance assessments in the approach proposed in this article are obtained implicitly, by interpreting a searcher's selection of one information object over others as an indication that the object is more relevant. The only assumption we make is that searchers will try to view information that relates to their needs; they will typically try to maximise the amount of relevant information they view whilst minimising the amount of irrelevant information (Pirolli & Card, 1995). The Ostensive Model (Campbell & Van Rijsbergen, 1996) is based on such principles and uses passive observational evidence, interpreted by the model, to adapt to searchers' current information needs. The rationale behind our work is similar to the Ostensive Model (i.e., we present an approach for approximating searcher interests through the implicit monitoring of their interaction) although the methodology is different. Our technique presents searchers with a variety of query-relevant document representations and uses searcher interaction with these and paths that join them as implicit RF. We weight terms using a *Binary Voting Model* (White, Jose, & Ruthven, 2003a) and choose the most useful terms selected by this model to modify the initial query.

To investigate the proposed approach of implicit RF we present an interface that we use to encourage interaction and hence generate more evidence for the techniques we employ. Studies by Spink et al. (1998) and White, Jose, and Ruthven (2003b, 2003c) showed that encouraging searchers to interact more with the results of their search can help them resolve vague information needs. Information needs may be dynamic and hence change in light of new information (Kuhlthau, 1993). Over time, we apply statistical methods to successive lists of potential query modification terms and use the resultant evidence to estimate the degree of change in a searcher's information need. As we show, different degrees of change result in different interface responses.

The interface offers two forms of support: the implicit selection of terms to expand the query and an estimation of the degree of development in the information need. We compare our implicit feedback approach with an experimental baseline, where the searcher is responsible for creating queries and making search decisions. In this comparison we evaluate how well our approach estimates information needs (through the terms it implicitly selects) and how well it estimates changes in these needs (through the interface support it offers).

The remainder of the article is structured as follows. In Section 2 we discuss the motivation behind our investigation and related work. Section 3 describes the document representations used and the search interface that uses them. The Binary Voting Model used for term selection is presented in Section 4 and the means of tracking need change in Section 5. In Section 6 we describe the evaluation of our approach. We present our results in Section 7, discuss them in Section 8 and conclude in Section 9.

2. Motivation

RF depends on a series of relevance assessments made *explicitly* by the searcher. The nature of the process is such that searchers must visit a number of documents and explicitly mark each as either relevant or non-relevant. This is a demanding and time-consuming task that can place an increased cognitive burden on those involved (Beaulieu & Jones, 1998; Morita & Shinoda, 1994). Searchers may be unwilling to provide relevance feedback and techniques that can estimate searcher feedback are therefore appealing.

Through implicitly monitoring interaction at the results interface, searchers are no longer *required* to assess the relevance of a number of documents, or indeed consider entire documents for relevance. Our approach makes inferences based on interaction and selects terms that approximate searcher needs.

Many studies have used a variety of 'surrogate' measures (hyperlinks clicked, mouseovers, scrollbar activity, etc.) (Lieberman, 1995; Joachims, Freitag, & Mitchell, 1997) to unobtrusively monitor user behaviour and estimate their interests. Through such means, similar studies estimate document relevance implicitly (Hill, Hollan, Wroblewski, & McCandless, 1992; Konstan et al., 1997; Nichols, 1997), and infer relevance from a variety of measures including the time spent viewing a document. If a user 'examines' (Kim, Oard, & Romanik, 2000) a document for a long time, or if a document suffers a lot of 'read wear' (Hill et al., 1992) it is assumed to be more relevant. However, these measures can be unreliable indicators of relevance (Kelly & Belkin, 2001) and also use interaction with the full-text of documents as implicit feedback. It is the relevant parts that contribute most to satisfying information needs and the remainder of the document may be erroneous, irrelevant or inappropriate. In our approach searchers can interact with different representations of each document. As will be described, these representations are of varying length, are focused on the query, have different origins and are logically connected at the interface to form an interactive search path.

Traditional RF systems require the searcher to instruct the system to perform RF, i.e., perform query modification and produce a new ranked list of documents (Salton & Buckley, 1990). However, this is only one way of using relevance information and may not always be appropriate. Information needs may be dynamic and can change in a dramatic or gradual manner (Bruce, 1994; Robins, 1997). For gradual changes, the generation of a new result set is perhaps too severe, and revisions that reflect the *degree* of change may be more suitable. We propose a form of *strategic help* (Brajnik, Mizzaro, Tasso, & Venuti, 2002) that makes decisions on how best to organise the search process and the information space to improve search effectiveness.

Our approach uses the evidence it gathers to track potential *changes* in information need and adopt a retrieval strategy that suits the *degree* of change. The underlying assumption is that large changes in estimated information needs result in new searches but smaller changes result in less radical operations, such as reordering the list of retrieved documents or reordering representations of the documents presented.

The main aim of our approach is to develop a means of better representing searcher needs whilst minimising the burden of explicitly reformulating queries or directly providing relevance information. Devising systems that adapt to the information needs of those who use them is an important step in developing systems to help searchers find the information they seek.

In the next section we describe the main interface components and the system that allows us to experiment on and evaluate the models that underlie our implicit feedback approach.

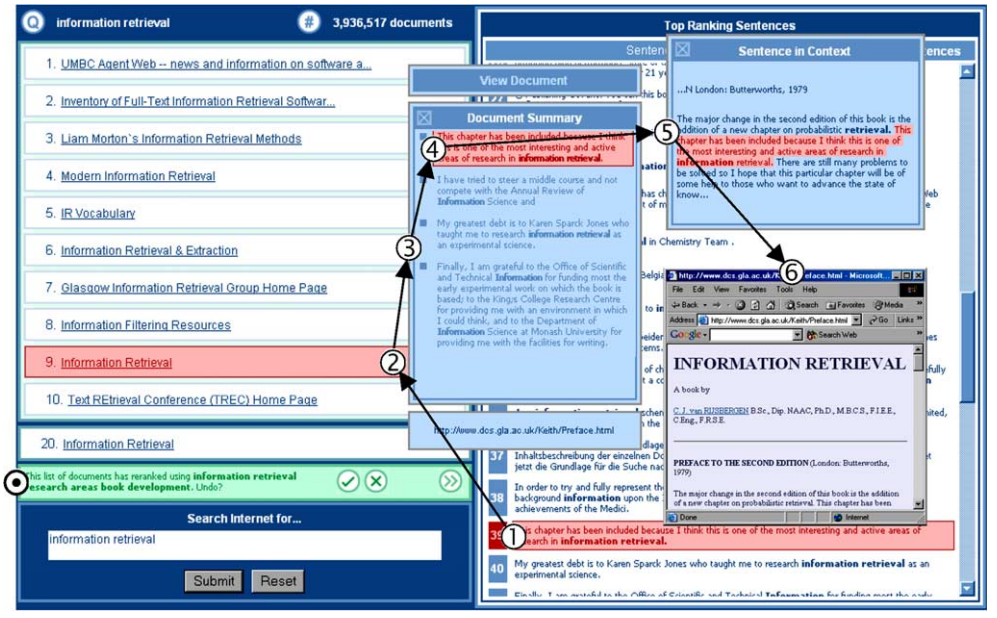

Fig. 1. Search interface with relevance path marked.

3. Searcher interaction

The approach described in this article gathers relevance information from searchers' exploration of the *information space*; the information content of the top-ranked retrieved document set. This space is created at retrieval time and is characterised by the presence of search terms (i.e., it is query-relevant). The space is re-created each time the document collection is re-searched. Exploring it allows searchers to examine search results more deeply and facilitates access to potentially useful information. Searchers can interact with *document representations* and follow *relevance paths* between these representations, generating evidence for our implicit feedback approach. In this section, we describe document representations and relevance paths and a search interface that combines them for interactive experimentation.

3.1. Document representations

The full-text of documents can contain irrelevant information. In our approach, we shift the focus of interaction to the query-relevant parts of documents and reduce the likelihood that erroneous terms will be selected by our approach. In Fig. 1, we present the search interface used in our experiments. This interface incorporates multiple document representations, and provides an example implementation of the models presented in this paper.

As well as being represented in the information space by their full-text (6) [1], documents are also represented by a number of smaller, query-relevant representations, created at retrieval time. These comprise the title (2) and a query-biased summary of the document (3) (White et al., 2003c). A list of sentences from the top thirty documents retrieved scored in relation to the query, called the *top-ranking sentences* (TRS), includes sentences from each document (1). Each of the document sentences included in the top-ranking

[1] The numbers correspond to those in Fig. 2.

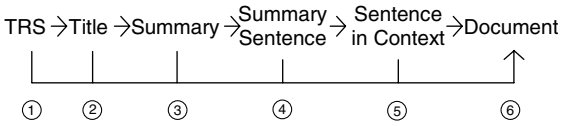

Fig. 2. The relevance path (numbers correspond to representations in Fig. 1).

sentence list is regarded as a representation of the document, as is each sentence in the summary (4). Finally, for each sentence in the summary there is a sentence in the context it occurs in the document (i.e., with the preceding and following sentence from the full-text of the document) (5). All document representations (except document titles) are sentence-based. Sentences are preferred to paragraphs (as used in passage retrieval; Callan, 1994; Salton, Allan, & Buckley, 1993) simply because they take less time to assess. This allows searchers to make speedy judgements on the relevance/irrelevance of the information presented to them.

3.2. Relevance paths

The six types of document representations combine to form a *relevance path* (shown in Fig. 2). The further along a path a searcher travels the more relevant the information in the path is assumed to be. The paths can vary in length from one to six representations and searchers can access the full-text of the document from any step in the path. Certain aspects of the path order are fixed e.g., the searcher must view a summary sentence before visiting that sentence in context.

As a searcher moves along the relevance path they move from assessing document representations in relation to other representations (i.e., top-ranking sentences, titles) to a deeper examination of representations in their resident context (i.e., summaries, sentences in context). Representations were arranged on the relevance path in such a way that the traversal of the path allowed subjects to progressively discover more information about the document.

Some representations of each document are fixed in content, i.e., the title and full-text of the document, whereas other representations, such as the summary, are dependent on the query and hence variable in content. Since path content is query-dependent, for each document there may be many *potential* relevance paths. We use the representations viewed by the searcher as evidence in our approach (Sections 4 and 5) and the distance travelled along the path as the strength of this evidence.

3.3. Search system

We developed a prototype system to experiment with our approach. The interface has an underlying functionality that allows it to connect to existing Web search engines. In this case we chose AlltheWeb [2] powered by the FAST search system to perform the Web searching. Since this search engine does not assume the default Boolean operator AND, it is more appropriate for query expansion.

Once the underlying search engine has performed a retrieval, the system downloads and summarises the top thirty ranked documents. Summarisation is carried out using a sentence extraction method described in White et al. (2003c). An inverted index of the top-ranked documents is also created. We use the vocabulary list from this index as the vocabulary in our Binary Voting Model described in the next section.

A searcher can access the document from any representation, so there is no need for them to follow the whole path to reach a document. The title is the default way of accessing a document, and the default

[2] http://www.alltheweb.com

results display shown after the initial retrieval and before any interaction is the list of titles and top-ranking sentences.

In Fig. 1 the searcher is looking for information on 'information retrieval' and is currently following a relevance path. The list of documents has just been reordered using the original query plus the top four-expansion terms 'research', 'areas', 'book' and 'development'. In the next section, we describe the Binary Voting Model that uses searcher interaction to choose such terms and retrieval strategies.

4. Binary Voting Model

In this section we describe the *Binary Voting Model*, a heuristic-based implicit feedback model developed to implicitly select terms for query modification. We use an approach that utilises searcher interaction with the document representations and relevance paths described in the previous section. The representations viewed by a searcher are used to select new query terms and in the Binary Voting Model each representation 'votes' for the terms it contains. When a term is present in a viewed representation it receives a 'vote', when it is not present it receives no vote. [3] All non-stopword, non-stemmed terms in the top-ranked documents are candidates in the voting process; these votes accumulate across all viewed representations. The assertion we make is that the winning terms are those with the most votes, and hence best describe the information viewed by the searcher. We assume that useful terms will be those contained in many of the representations that the searcher chooses to view.

However, different types of representation vary in length, and can hence be regarded as being more or less *indicative* of the content of the document (Barry, 1998). Representations with a higher indicativity are regarded as providing better quality evidence for the implicit feedback approach. For example, a top-ranking sentence is less indicative than a query-biased document summary (typically composed of four sentences) as it contains less information about the content of the document. To counter this we *weight* the contribution of a representation's vote based on the indicative worth of the representations, e.g., we consider the contribution that viewing a top ranking sentence makes to the system computing which terms are relevant to be less than a summary.

The weights used in our experiments are 0.1 for title, 0.2 for TRS, 0.3 for Summary, 0.2 for Summary Sentence and 0.2 for Sentence in Context. For example, all terms in a viewed summary will receive a weight of 0.3, all terms in a viewed summary sentence will receive a weight 0.2, etc. These weights were defined for experimental purposes and were based on the typical *length* of a representation. They are used a heuristic measure of representation quality, ensure that the total score for a term is between 0 and 1 (inclusive) and are used in the absence of a more formal methodology.

The Binary Voting Model is a simple approach to a potentially complex problem. The terms with most votes are those that are taken to best describe the information viewed by the searcher (i.e., those terms that are present most often across all viewed representations) and can therefore be used to approximate searcher interests.

As shown in Fig. 2, multiple representations can form a relevance path for each document. We use the distance travelled along the path and the particular representations used in the path to calculate a list of expansion terms for query modification.

Each document is represented by a vector of length n, where n is the total number of unique non-stopword terms in the top 30 Web documents. [4] We refer to the list holding these terms as the *vocabulary*.

[3] The motivation and experimental justification for using binary presence/absence information is given in White et al. (2003a).

[4] Only 30 retrieved documents are used for analysis to ensure the system responds in a timely manner.

$$\begin{array}{c} \\ Q_0 \\ D_1 \\ D_2 \\ \ldots \\ D_d \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \ldots & t_n \end{array} \\ \begin{bmatrix} t_{01} & t_{02} & \ldots & t_{0n} \\ t_{11} & t_{12} & \ldots & t_{1n} \\ t_{21} & t_{22} & \ldots & t_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ t_{d1} & t_{d2} & \ldots & t_{dn} \end{bmatrix} \end{array}$$

Fig. 3. Document × term matrix.

To weight our terms we build a document × term matrix, $(d + 1) \times n$, where d is the number of documents for which the searcher has travelled at least part of the path (Fig. 3). Each row in the matrix is all n terms in the vocabulary [i.e., $(t_{k1}, t_{k2}, \ldots, t_{kn})$ where k is the row number], and each term has a weight. An additional row is included for the query.

Query terms are initially assigned a weight of 1 if they are included in the query and 0 if not. Example 1 (used throughout this section) illustrates how the Binary Voting Model operates.

Example 1. *Simple updating*

If we assume that there are only 10 terms in the vocabulary and that the original query (Q_0) contains t_5 and t_9:

$$\begin{array}{c} \\ Q_0 \end{array} \begin{array}{c} \begin{array}{cccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

This vector is then normalised to give each term a value in the range [0, 1] and make the values sum to one. This ensures that the query terms are not weighted too highly in the document × term matrix. The vector now looks like:

$$\begin{array}{c} \\ Q_0 \end{array} \begin{array}{c} \begin{array}{cccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 \end{bmatrix} \end{array}$$

We treat each document representation as a source of terms, and the act of viewing a representation as an implicit indication of relevance. When a searcher visits the first representation for a document we add a new row to the document × term matrix. This row is a vector of length n, where n is the size of the vocabulary and all entries are initially set to 0. If a term occurs in a representation, no matter how many times, it is assigned a weight, w_t, which is based on the representation that contains the term.

This weight for each term is *added* to the appropriate term/document entry in the matrix. Weighting terms is therefore a *cumulative* process; the weights calculated for a term in one representation are added to the weights calculated for the preceding steps in the relevance path. Unlike standard RF algorithms which calculate one set of weights for expansion terms, our system calculates weights on a per document basis. That is, we have different sets of weights for each document.

The total score for a term in a document is computed by

$$w_{t,D} = \sum_{j=1}^{p} (w_{t,r}) \tag{1}$$

where p is the number of steps taken, D is the document, t is the term, r is the representation and $w_{t,r}$ is the weight of t for representation r.

Example 1. (continued) *Simple updating*

$$\begin{array}{c} \\ Q_0 \\ D_1 \\ D_4 \end{array} \begin{array}{c} \begin{array}{cccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 \\ .4 & 0 & 0 & .1 & .4 & 0 & .2 & .2 & .7 & 0 \\ .1 & .1 & .2 & 0 & .2 & 0 & 0 & 0 & .1 & 0 \end{bmatrix} \end{array}$$

If document × term matrix is in this state and the searcher expresses interest in the title of document D_1—which has a step weight of 0.1, and contains terms t_1, t_2 and t_7—the matrix changes to:

$$\begin{array}{c} \\ Q_0 \\ D_1 \\ D_4 \end{array} \begin{array}{c} \begin{array}{cccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 \\ \mathbf{.5} & \mathbf{.1} & 0 & .1 & .4 & 0 & \mathbf{.3} & .2 & .7 & 0 \\ .1 & .1 & .2 & 0 & .2 & 0 & 0 & 0 & .1 & 0 \end{bmatrix} \end{array}$$

The weights of terms t_1, t_2 and t_7 are directly updated. The term t_2 is now seen as being important to document D_1 (before its weight was 0). Terms t_1 and t_7 are seen as being *more* important than before to D_1 (their previous weights were 0.4 and 0.2 respectively).

If the searcher visits one representation of a document and then goes onto the next representation in the path of that document, *at any time—not necessarily immediately*, we add the term scores to the row in the matrix occupied by that document. The scoring is cumulative; if a document already has a row in the matrix it does not get a new one. Since indicativity weights sum to one, the value in any cell in the matrix cannot exceed one.

Similarly, if the searcher views the same representation twice, e.g., the same summary twice, the model only counts the representation once. In effect, it keeps a history of which representations have been viewed. The current version of our approach does not consider more detailed interaction. We cannot differentiate, for example, between a searcher seeking relevant information and a searcher checking what they have already examined, something that may account for them looking at the same representation twice.

The matrix resulting from this process reflects the weights based on all paths viewed by the searcher. This information is used for query modification, as will be described in the next section.

5. Information need tracking

To provide an appropriate level of support to the searcher, our approach uses a history of recent interaction and predicts changes in the information need. This history provides insight into the recent interests of the searcher, and by comparing this with previous histories we track possible changes in the information need. Selecting the most appropriate form of support depends on the extent to which the need is seen to change. The smaller the change, the less radical the support offered. Tailoring the support in this way allows the interface to work in concert with the searcher.

In the matrix created by the Binary Voting Model, only the query terms and terms in representations viewed by the searcher will have a score greater than zero. The latter set of terms is potentially useful for query modification. One novel aspect of our system is how we use these terms. The approach detects changes in the set of terms it suggests for query modification and, based on the degree of change, computes how the new query should be used. In this section, we describe how the new query is formed and how this query is used.

5.1. Query creation

For every five paths the approach computes a new query. This allows the approach to gather sufficient implicit evidence from searcher interaction. It is possible for a relevance path to contain only one represen-

tation. Therefore, for the searcher to follow five paths they need only view five unique document representations. To compute the new query we calculate the average score for each term across all documents (i.e., down each column in the matrix). This gives us an average score for each term in the vocabulary. The terms are then ranked by this score. A high average score implies the term has appeared in many viewed representations and/or in those with high indicative weights across the documents viewed.

The top six ranked terms are used to form the new query. This number of terms had been shown to be effective in our earlier work (White, Ruthven, & Jose, 2002a, 2002b) and in related work (Harman, 1988).

Example 1. (continued) *Simple updating*

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Q_0	0	0	0	0	.5	0	0	0	.5	0
D_1	**.5**	**.1**	0	.1	.4	0	**.3**	.2	.7	0
D_4	.1	.1	.2	0	.2	0	0	0	.1	0
average score	.2	.07	.07	.03	.37	0	.1	.07	.4	0

The top six terms chosen from this matrix are: t_9, t_5, t_1, t_7, t_3 and t_2. The query terms received multiple 'votes' and are ranked highest by the Binary Voting Model. Although terms t_2, t_3 and t_8 have the same score, t_8 is not included since t_3 occurs more recently (in document D_4) and t_2 occurs in more than one document.

It is possible that the new query may not contain the searcher's original query terms; this would be a form of query replacement as the estimated information need has changed sufficiently to warrant the original query being completely replaced.

5.2. Query application

For each set of thirty retrieved documents the vocabulary [5] is static, so we can gauge the level of change in the information need by computing the change in the term ordering from the term list at step m (i.e., q_m) and the list at step $m + 1$ (i.e., q_{m+1}). As the vocabulary is static, the *terms* in the list will not change, only their order. So, by comparing q_m against q_{m+1} based on some operator $\circ$ we can compute the degree of change between the lists and therefore predict changes in the information need. This can be shown formally as

$$\Delta\psi = (q_m) \circ (q_{m+1}) \tag{2}$$

where ψ is the system's view of the searcher's information need and $\circ$ computes the difference between two lists of unique terms. In our approach, we use the *Spearman rank-order correlation coefficient* as the operator $\circ$. The correlation returns values between -1 and 1 and is non-parametric, so rankings, not the actual term scores, are used. A correlation of 0 implies *zero* (or no) *correlation* between the two lists. We can calculate $\Delta\psi$ using the coefficient as follows:

$$\Delta\psi = \frac{\sum_{i=1}^{n} r(q_{m_i}) r(q_{m+1_i}) - n\left(\frac{(n+1)}{2}\right)^2}{\left(\sum_{i=1}^{n} r(q_{m_i})^2 - n\left(\frac{(n+1)}{2}\right)^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} r(q_{m+1_i})^2 - n\left(\frac{(n+1)}{2}\right)^2\right)^{\frac{1}{2}}} \tag{3}$$

[5] The list of all unique, non-stemmed, non-stopword terms present in the top 30 retrieved documents.

where we assume that q_m and q_{m+1} are both ranked lists of terms, $r(\cdot)$ is the rank of a term from a list and n is the total number of terms. We handle ties in the standard statistical way, by summing the rank of all tied elements and dividing this sum by the number of elements, effectively taking the average rank for each group of ties.

All terms in the original vocabulary are ranked based on the weights derived from the Binary Voting Model, and averaged across all viewed documents. These terms are present in both lists (q_m and q_{m+1}) but potentially in a different order, depending on the representations viewed by the searcher. There is a high level of redundancy in each list as the lower ranking terms that never appear in a viewed representation experience only slight changes in their ranking between iterations. To counter this problem we use only the top 100 terms in our calculations. The top 100 terms were used since these are the most liable to change and hence most likely to reflect any change in the information viewed. As the number of terms increases (i.e., greater than 100), redundancy in the term list also increases and the predicted level of change becomes more conservative. In contrast, as the number drops (i.e., less than 100) the likelihood of change increases, making the prediction more radical.

We compare the lists every time we create a new query (i.e., every five relevance paths). To compute the correlation coefficient both lists must contain the same terms and the same number of terms. Therefore, in practice we need to use the first 100 terms plus β, where β is the number of terms that have left or joined the top 100 terms between q_m and q_{m+1}. For terms *joining* the top 100, we sort them based on their original (q_m) ranks and assign them ranks (in q_m) in the range $[101, 101 + \beta]$. We use the same procedure for terms that are *leaving* the top 100, except these terms are ranked based on their new (q_{m+1}) ranks (Fig. 4).

We then have the coefficient in the range $[-1, 1]$, where a result closer to -1 means the term lists are dissimilar with respect to their rank ordering. As the coefficient gets closer to 1, the similarity between the ranking of the terms in the two query lists increases. Based on the coefficient value returned we decide how to use the new list of terms. Fig. 5 shows the boundaries of the Spearman correlation coefficient that are used to select retrieval strategies.

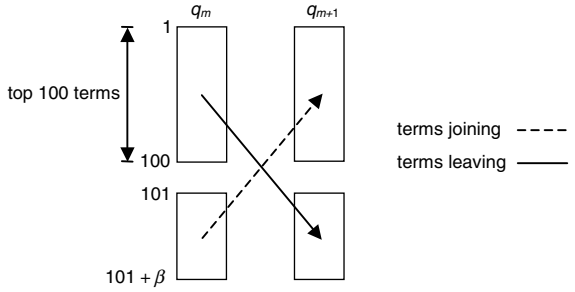

Fig. 4. Terms leaving and joining the first 100.

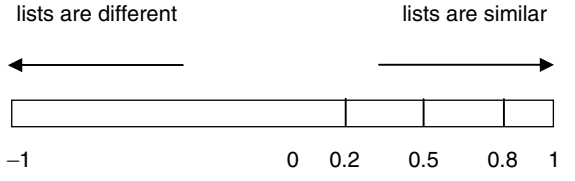

Fig. 5. Decision boundaries of the information need tracking component.

We implement three strategies:

i. *Re-searching*. If the coefficient value indicates that the two term lists are substantially different with respect to rank ordering, we take this to reflect a large change in ψ (the information need as identified by the approach). In this situation we carry out a re-search to retrieve a new set of document. A new result set is generated in the background and the searcher is informed by a message at the interface. The searcher must request to view the results of the new search; the new documents are not automatically shown. Coefficient values of less than 0.2 are taken to indicate a large change in the term lists.
ii. *Reordering documents*. A result in the range $[0.2, 0.5)$ indicates a weak correlation between the two lists and consequently a less substantial change in ψ. Here we use the new query (i.e., the six top-ranked terms) to reorder the top 30 retrieved documents. We reorder the document list using best-match *tf.idf* scoring with the expanded query. The vocabulary list remains unchanged after this action.
iii. *Reordering TRS*. Coefficients in the range $[0.5, 0.8)$ indicate a strong correlation between the two term lists and hence a small change in the predicted of the information need. In this case we use the new query to re-rank the TRS list. The sentences are the most granular elements presented to the searcher and are therefore most suited to reflect minor changes in ψ. The top ranking sentences are reordered based on the term-occurrence of each of these expansion terms.

Strategies ii and iii provide an updated view of the retrieved documents based on the current ψ. For differences between 0.8 and 1, the need is assumed to have not changed sufficiently to warrant action. All numerical bounds are arbitrary, chosen during pilot testing of the implicit RF approach.

In the implicit RF system that implements the approach the retrieval strategy occurs automatically. The system notifies the searcher with a message at the periphery of the interface (shown by $\odot$ in Fig. 1) and highlights the part of the interface affected by the action. The message includes the generated query and gives the searcher the option to reverse the action's effect. The system acts on the searcher's behalf, *then* offers them the option to reverse the action. An alternate strategy could be to offer searchers control over when the action occurred. We felt that it was better to show searchers the output of the action and let them decide on the value of the action, rather than let them rely on the *perceived value of the potential action*, a judgment that may be tainted by previous experience. There is no need for the searcher to respond to this message, and it will disappear after a short time. In the next section, we describe a user-centred evaluation of our approach.

6. Evaluation

We evaluated the ability of our approach to track information needs using human subjects and different search scenarios. We use two systems; one implementing the implicit feedback approach described thus far and a manual baseline system that placed responsibility for query reformulation and action on the searcher. The baseline system is described in more detail in Section 6.4.

We specifically tested three main hypotheses:

Hypothesis 1. The terms selected for implicit feedback represent the information needs of the subject (i.e., term selection support).

Hypothesis 2. The implicit feedback approach estimates changes in the subject's information need.

Hypothesis 3. The implicit feedback approach makes search decisions that correspond closely with those of the subject.

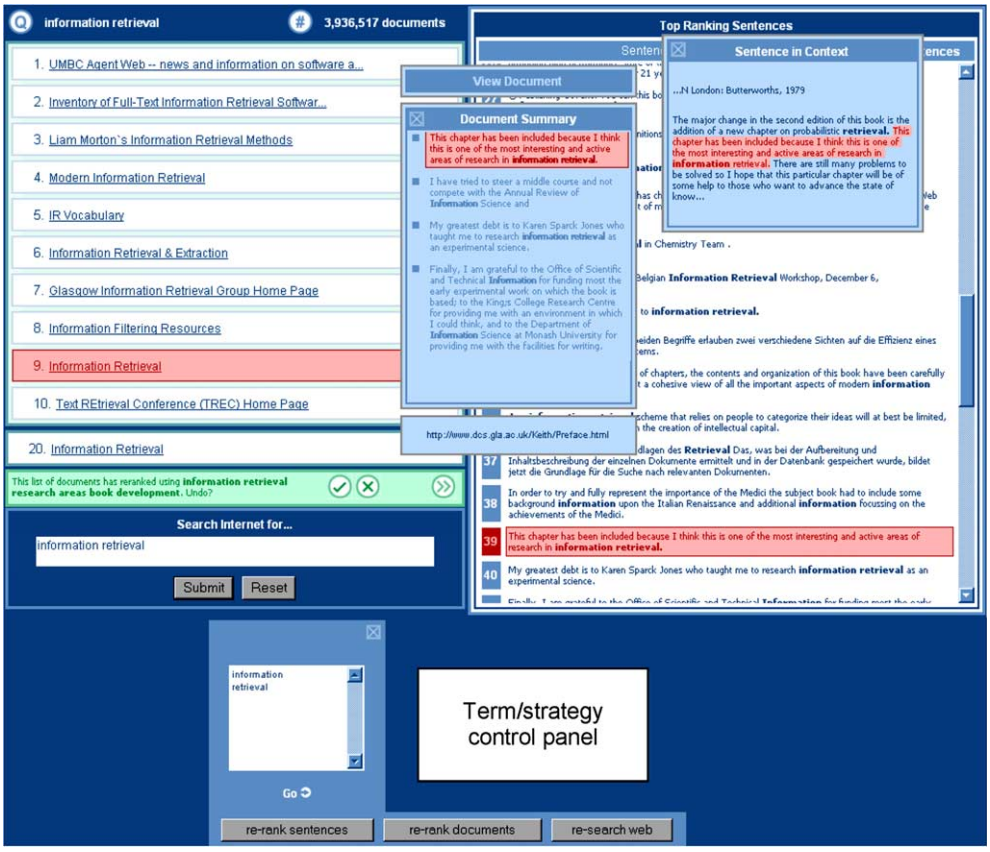

Fig. 6. Baseline search interface.

These hypotheses test two aspects of our approach; the Binary Voting Model and the estimation of information need change. In this section, we provide details of our experimental subjects, the tasks, the experimental methodology employed and the baseline system.

6.1. Baseline system

The baseline system uses the same interface components as the implicit feedback system but differs in one key way; the searcher is solely responsible for adding new query terms and selecting what action [6] is undertaken after these terms have been added. These options give the searcher increased control over the search but also increased responsibility for making decisions.

The baseline interface (Fig. 6) contains one additional component; a term/strategy control panel. This panel allows searchers to decide how best to use the query. Subjects were informed that each of the three strategies provided a different (and increasing) level of interface support. To invoke one of the three possible strategies, the searcher must click one of the three buttons shown in Fig. 6. A small window then ap-

[6] Reordering top-ranking sentences, reordering documents or re-searching Web.

pears directly above the button clicked. In Fig. 6 the *re-rank sentences* option has just been clicked and the original query terms 'information' and 'retrieval' are shown in the scrollable window.

Searchers make all decisions about query modification and can therefore expand and, if required, replace their original query. This is similar to the implicit feedback system, which will replace original query terms if the estimated level of information need change is sufficiently large. Once satisfied with their modified query, searchers can instruct the system and the selected action is executed using the newly formulated query.

The nature of this baseline allowed us to evaluate how well the implicit feedback system detected information needs *from the perspective of the subject*. We tested whether the approach chose terms and actions that matched those chosen by the subject and whether the subject felt the support offered was beneficial. Systems that use implicit feedback techniques are potentially unpopular since they remove searcher burden but *also* searcher control. In this study, we acknowledge this, and compare the approach with a baseline where subjects has such control. Since the aim of our approach is to estimate changes in information needs *and not to propose an alternative to explicit relevance feedback*, we reserve a comparison with explicit RF for future work and regard similar inter-system results (i.e., where the implicit feedback system performs as well as the baseline) as promising.

6.2. Experimental subjects

We recruited 24 subjects for our experiments. Subjects were mainly undergraduate and postgraduate students at the University of Glasgow and were recruited using electronic mails, poster advertisements and word of mouth. Our recruitment was specifically aimed at targeting two groups of subjects: *inexperienced* and *experienced*. The experienced subjects were those who used computers and searched the Web on a regular basis. Inexperienced subjects were those who searched the Web, used computers and the Internet infrequently. On average per week, inexperienced subjects spent 3.1 h online, and experienced subjects spent 34.9 h online. Experienced subjects were generally studying for a Computing Science related degree and inexperienced subjects were generally studying for degrees in Arts and Social Sciences. Overall, our subjects had an average age of 26 with a range of 38 years (youngest 16 years, oldest 54 years). 14 males and 10 females participated in the experiments.

The classification between experienced and inexperienced subjects was made on the basis of the subjects' responses to questions about their experience and their own opinion of their skill level.

6.3. Experimental tasks

Each subject was asked to complete one search task from each of four categories, each containing two tasks. An example task is included in Appendix A and all other tasks in White (2004). The categories were: *fact* search (e.g., finding a named person's current e-mail address), *decision* search (e.g., choosing the *best* financial instrument), *background* search (e.g., finding information on dust allergies) and *search for a number of items* (e.g., finding contact details for a number of potential employers) (White et al., 2002a). Each search task was placed within a simulated work task situation (Borlund, 2000). The proponents of this technique assert that subjects should be given search scenarios that reflect real-life search situations and should allow the subject to make personal assessments on what constitutes relevant material. The different tasks engender realistic search behaviour and produce different types of simulated information needs within the range of verificative and conscious topical information needs (Ingwersen, 1992).

There were two tasks per category, each of a similar level of difficulty (verified by *a priori* pilot testing and questions in the post-task questionnaire) and subjects were asked to choose the task they would like to do. Subjects chose 51.0% of tasks because they were *interesting*, 21.8% of tasks because they felt they were *easy*, 19.8% because they were *familiar* with the topic area and 7.4% for *no reason*. Whilst the subject groups

were homogeneous (i.e., inexperienced or experienced) no criteria other than search experience were used in the selection of candidates. Subject interests were potentially diverse and it was not possible to offer a single task in each task category that appealed to all subjects. Giving subjects a choice of tasks in each category increased the likelihood that the task they chose would interest them. This supports the work of Borlund (2003), who suggests that interest in the topic of the task is an important factor in the design of simulated situations.

Offering subjects a choice of tasks allowed them to choose tasks that interested them and were familiar. Subjects chose one search task from two for each task category. Subjects with topic experience are better equipped to make expansion decisions using that topic's terms and relevance assessments of that topic's documents (Vakkari, 2002).

6.4. Experimental methodology

In our experiments 24 subjects completed four search tasks, two tasks on each of the two systems: implicit feedback and baseline. The presentation of tasks to subjects was held constant; each subject performed the search tasks in the same order, however the order of presentation of systems was rotated across subjects. The tasks had been used in previous experiments (White et al., 2003c), where the impact of task bias was not significant. Subjects were given a maximum of 10 min to complete each task. These time constraints allowed different systems and search tasks to be compared fairly since subjects were given the same amount of time on each system. The Web was used as the collection for these experiments since subjects had experience interacting with Web documents, effective retrieval systems were readily available and realistic search scenarios could easily be created.

The subjects were given a short tutorial on the features that were incorporated into the two systems being tested and a training task to allow them to become accustomed to both systems. We also collected background data on aspects such as the subjects' experience and training in online searching. After this, subjects were introduced to tasks and systems according to the experimental design. Subjects were instructed to attempt the task to the best of their ability and write their answer on a sheet provided. Since it may impact on how they used the system, subjects were not told *how* the Binary Voting Model and information need tracking worked. A search was seen to be successful if the subject felt they had succeeded in their performance of the task. This is closely related to real information seeking situations, where the goal of any retrieval system is typically to satisfy the searcher.

Once they had completed a search, subjects were asked to complete questionnaires regarding various aspects of the search. We used semantic differentials, Likert scales and open-ended questions to collect these data. All results from questionnaires were measured on a 5-point scale, where a rating closer to 1 corresponds to a stronger agreement. These forms of capturing subjective information have been effective in related work (Brajnik, Mizzaro, & Tasso, 1996). In addition, we conducted semi-structured interviews after each search and after the experiment as a whole. Background logging was used to record each subject-system interaction event (e.g., queries submitted, mouse clicks, etc.) with an associated time stamp.

In the next section we present and analyse our experimental results.

7. Results and analysis

We focus on results pertinent to the three research hypotheses; the terms chosen approximate the information needs of the subject, changes in information needs are estimated and retrieval strategies chosen are appropriate. In this section, we also present results on the novel interface components (i.e., the relevance paths and increased information content). We used simulated work tasks *only* to facilitate interaction with the interface and therefore do not consider measures of task success or effectiveness. We use independent

t-tests for comparisons *between subject groups* and *between systems*, and paired *t*-tests for *within-group system* [7] comparisons. We test the significance of our results at $p < .05$, unless otherwise stated. M is used in this section to denote the mean, and S_i and S_b to denote the implicit RF system and the baseline experimental systems respectively.

7.1. Hypothesis 1: Information need detection

We measured the value of the modified query constructed by the implicit feedback approach using the degree of overlap with terms chosen by the subject and subject opinion on their usefulness.

7.1.1. Term overlap

The implicit feedback system uses the Binary Voting Model to choose terms for query modification. In the baseline system the subject is responsible for making such query reformulation decisions without suggestions from the system.

We measure the degree of *term overlap* using the baseline system. In the baseline system, the Binary Voting Model runs in the background, completely invisible to the subject and not involved directly in any query modification decisions. That is, whilst the Binary Voting Model chooses terms based on subject interaction, these terms are never shown to the subject and never used to construct the new query. At any point in time, the model holds six terms that the would be used for query modification if the system was responsible. We measure the degree of *term overlap* based on how frequently terms chosen by the subject occur amongst the non-query [8] terms in these top six. High values of term overlap suggest that the terms chosen by the Binary Voting Model are of good value and match the subject's own impression of their information need. An example taken from the experiments is shown in Fig. 7, where the initial query (at iteration 0) is *dust allergies* and terms added by the subject that co-occur with system terms are shown in bold.

When terms chosen by the subject that are not in original query co-occur with terms chosen by our approach, regardless of the number of times, we regard that as an instance of term overlap. Table 1 shows the average percentage of occasions where the top six terms chosen by our approach also included as at least one of the subject's terms. This is shown for each task type and across all task types. The figures do not include occasions where subjects' original query terms co-occurred with system terms.

The table shows that on a high proportion of occasions any of the top six terms chosen by the implicit model co-occur with subjects' chosen terms. The difference between the inexperienced and experienced subjects was not significant. [9] We ran a one-way repeated measures ANOVA to test the significance of inter-task category differences within each subject group; no differences were significant. [10] However, the term overlap for experienced subjects was generally higher than that for inexperienced subjects. These differences may be attributable to: (i) *subject term selection* (i.e., differences in which terms the subject groups regarded as important) and (ii) *subject interaction* (i.e., differences between the subject groups in the type and amount of interaction). In the next section, we analyse query modification behaviour.

7.1.2. Query modification behaviour

In Table 2, for each system and per task category, we show the average number of query iterations and the average query length (not including the original query terms). In Table 2, an 'iteration' is regarded as the use of a query for any action; reordering the top-ranking sentences, reordering the documents or re-searching the Web.

---

[7] E.g., experienced searchers on implicit feedback system vs. experienced searchers on explicit system.

[8] Terms that are not included in the *original* query.

[9] $t_{22} = 1.44$, $p = .0819$.

[10] *Inexperienced*: $F_{3,95} = 2.14$, $p = .1003$; *experienced*: $F_{3,95} = 2.31$, $p = .0812$.

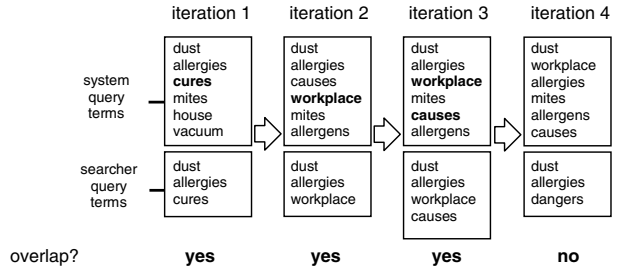

Fig. 7. Term overlap in query modification iterations.

Table 1
Term overlap occurrence (values are percentages)

Task type	Subject group	
	Inexperienced	Experienced
Fact	69.13	73.19
Decision	72.11	75.63
Background	70.24	74.62
Number of items	67.90	74.33
All	69.85	74.44

Table 2
Average number of query 'iterations' and average query length

	Task category				Significance at $F_{3,95}$
	Fact	Decision	Background	Number of items	
Baseline					
Average number of query 'iterations'	7.25	5.50	5.35	5.70	3.86 (p = .012, α = .025)
Average query length	1.45	2.30	3.00	2.75	3.14 (p = .029, α = .025)
Implicit					
Average number of query 'iterations'	4.55	5.00	4.75	6.45	4.65 (p = .005, α = .025)
Average query length	4.33	3.25	3.20	3.00	2.96 (p = .036, α = .025)

In the implicit feedback system the average query length is the number of terms in the new query that were not in the original query. The *search for a number of items* task was a variation of the fact search, requiring subjects to find more than one instance of relevant information.

We ran an ANOVA on each dependent variable. To reduce the number of Type I errors i.e., rejecting null hypotheses that were set to true, we set the *alpha level* (α) to .025 i.e., .05 divided by 2, the number of dependent variables. In the final column of Table 2 we show the statistical significance of the variation between variable values for each type of task. The results and further pair-wise comparisons between task types show that the *fact* search encourages significantly more reordering and re-searching, and significantly shorter queries. When a subject's need is well-defined they are more able to assess relevance, submit more queries and be more efficient at assessing results.

For each task category we compared the average number of query iterations between system types. In the fact search the system did not act as often as the subject, and in the search for a number of items it

Table 3
Query manipulation frequency in the baseline system

Manipulation activity	Task category				Significance at $F_{3,95}$
	Fact	Decision	Background	Number of items	
Adding	2.80	4.30	4.80	3.67	2.84 ($p = .042$, $\alpha = .025$)
Removing	2.05	1.33	1.40	2.55	.97 ($p = .410$, $\alpha = .025$)

acted more often; these differences were significant. [11] However, for the decision search and background search the frequency of implicit feedback system action was more in line with the baseline; these differences were not significant. [12] This suggests that the system acts in a way suited for when the work task undertaken is more uncertain.

The baseline system is the only system where the subject can directly manipulate the query. There are two types of query manipulation: adding terms and removing terms. The addition or removal of terms or a set of terms then using this revised query in some way (i.e., to perform a retrieval strategy) is an instance of a query manipulation activity. In Table 3 we show the average frequency of these activities for each subject performing different types of search.

The results show that on average, subjects typically added terms to their queries more often for the *decision* and *background* searches than for the *fact* search and *search for a number of items*. This does not suggest that they added a greater number of terms in total, simply that they made more decisions to add terms. The same terms could quite feasibly occur in many manipulation decisions and more than one term could be added for each decision. Pilot tests carried out before the experiment went some way to ensuring the task categories were of similar levels of difficulty. In each of the four task categories, no task was chosen more than the alternative.

7.1.3. Subject opinion

Participants were asked to rate (using a 5-point semantic differential) whether the terms added to their original query were useful *always, occasionally* or *never*. There were no significant differences in the comparison between systems ($M_b = 1.84$ vs. $M_i = 2.11$) [13] or between groups ($M_E = 1.90$ vs. $M_I = 2.01$). [14] However, it is worth noting that even though the terms selected by subjects were useful, those selected by the implicit feedback system were also useful and significantly less than the median rating (i.e., less than 3). [15] This is a positive result since subjects had no direct control over which terms were selected.

The term selection in the implicit feedback system was generally well received. From subject comments in informal interviews we can conjecture that it may be of most use when their information need is ill-defined and variable (i.e., in the decision search), and they need support from the system. When the need is well-defined and static (i.e., in the fact search) they had an exact idea of what they are searching for, and the implicit term selection may not be of as much use.

7.2. Hypothesis 2: Information need tracking

The implicit feedback system translated the results of the information need tracking component into differing degrees of interface support (i.e., reordering top-ranking sentences, reordering documents or

[11] *Fact search*: paired $t_{22} = 3.55$, $p = .0009$, *search for a number of items*: paired $t_{22} = 2.62$, $p = .0078$.
[12] *Decision search*: paired $t_{22} = 1.26$, $p = .1104$, *background search*: paired $t_{22} = 1.18$, $p = .1253$.
[13] $t_{22} = 1.38$, $p = .0907$.
[14] $t_{22} = 1.65$, $p = .0566$.
[15] $t_{22} = 2.86$, $p = .0046$.

Table 4
Subject perceptions of the search strategy

Likert scale	Subject group			
	Inexperienced		Experienced	
	S_b	S_i	S_b	S_i
Occurred at appropriate time	1.31	1.54	1.48	1.50
Did not annoy subject	1.65	1.82	2.10	2.29
Helpful for completing the task	2.35	2.47	1.97	2.28

researching the Web). Of the three interface support options, subjects would perhaps be more familiar with using a revised query to re-search the Web. Therefore, we measure how well the approach tracked information needs through how subjects responded to the decisions the implicit feedback system made on their behalf. We measured the value of these search *tactics* (Bates, 1990) by eliciting subject opinion and studying subject interaction.

7.2.1. Subject perceptions

In the post-search questionnaires, completed after each search task, subjects were asked to provide feedback on how they perceived the search strategy [16] used in each experimental system. In the implicit feedback system this was the automatic reordering or re-searching, and in the baseline this was controlled by the subject. The implicit feedback system acted every five paths. However, since there are many potential paths for a document this does not imply that the subject must view five unique relevance paths from the system to act. The result presentation techniques used encourage subjects to interact more with the results of their search, and if required restructure the available information (reorder documents or top-ranking sentences). We discourage search strategies that involve multiple query submissions and brief examinations of only the first page of search results (Jansen, Spink, & Saracevic, 2000).

Subjects were asked to respond using 5-point Likert scales, where a lower value reflects an increased level of agreement. For example, the differential for the first entry in Table 4 was *the action occurred at an appropriate time*.

In the baseline system, the subject has control of the search strategy employed, and we would expect their responses to be more positive than the implicit feedback system. This is the case; however, what is interesting is that the differences between the systems are not statistically significant with MANOVA across all differentials. [17] A two-way repeated measures ANOVA was run on each of the differentials. The results showed no significant inter-system differences, but a significant inter-group difference for the differential *the action did not annoy the subject*. [18] However the difference between M_b and M_i for the differential *the action was helpful for completing the task* for experienced subjects seemed large and warranted further investigation. We ran a paired *t*-test between the ratings obtained in this differential (i.e., experienced subjects on baseline vs. experienced subjects on implicit feedback system). The results were not significant. [19] From questionnaire responses and informal interviews we conjecture that the search strategy employed by the implicit feedback system was an approximation of subjects' own intentions. This is a promising result, since in a study of this nature we would expect subjects to perceive their own choice of strategy as better than the system's.

[16] Also referred to in this article as the 'action'.
[17] $F_{3,42} = 2.01$, $p = .1271$.
[18] $t_{22} = 2.78$, $p = .0055$.
[19] $t_{22} = 1.65$, $p = .0566$.

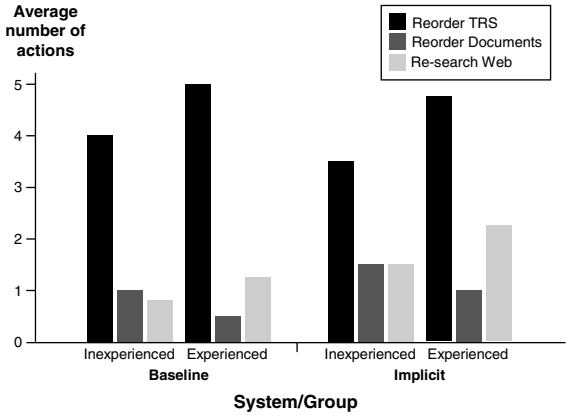

Fig. 8. Average number of actions performed.

Implicit feedback systems can be unpopular since they remove direct control for actions, such as query formulation, that subjects may be familiar with. Using 5-point semantic differentials where a lower score is better, we tested the extent to which they felt in control of their search ($M_b = 1.24$ vs. $M_i = 1.43$) and that the system intruded on their search ($M_b = 1.65$ vs. $M_i = 1.79$ [20]). The implicit feedback does not have a detrimental effect on the search process.

7.2.2. Search strategy

Fig. 8 shows the average number of actions carried out on each system across all search tasks. For both subject groups and both systems, the results show significant differences (with a paired t-test, t_{22}) in the number of times the top-ranking sentences were reordered, compared to other actions. [21] There were no significant inter-group differences (i.e., all $p \geqslant .19$), however experienced subjects tended to make more use of the unfamiliar actions, i.e., the reordering of the top-ranking sentences and the reordering of the top-ranked retrieved documents, than the inexperienced subjects.

When given control over the action performed, both groups of subjects chose to reorder the list of top-ranking sentences more than the implicit feedback system and reorder the documents less frequently. Experienced subjects re-searched the Web and reordered the top-ranking sentences, but chose to reorder the documents less frequently. Further explanatory evidence, such as the rationale behind this behaviour, was not captured in this study. Questionnaires and post-search interviews may have provided more insight into the reasoning behind their actions, although this is dependent on subjects being able to provide them.

Studies of search behaviour (Chen & Dhar, 1991; Jansen et al., 2000) have shown that subjects prefer to use the trial-and-error strategy of query reformulation and resubmission, rather than deeply examining the result set content. Subjects are familiar with such a strategy and it can be effective if they understand their information need. However, in White et al. (2003c) and Spink et al. (1998) it was demonstrated that a deeper examination of the retrieved document set can lead to more effective searching, especially in circumstances where the information need is vague. The reordering of sentences and documents allows the implicit feedback system (or the subject in the baseline system) to reshape the information space to suit their current information needs. This facilitates more interaction with the contents of the retrieved document set, and lessens the need to submit a number of queries.

[20] In control ($t_{22} = 1.32$, $p = .1002$), intruded ($t_{22} = .98$, $p = .1689$).

[21] *inexp/baseline*: TRS/doc ($p < .0002$), TRS/Web ($p < .0002$); *exp/baseline*: TRS/doc ($p < .0001$), TRS/Web ($p < .0002$); *inexp/implicit*: TRS/doc ($p < .0003$), TRS/Web ($p < .0003$); *exp/implicit*: TRS/doc ($p < .0001$), TRS/Web ($p < .0002$).

Table 5
Percentage of reversed actions per search session

Retrieval strategy	Subject group			
	Inexperienced		Experienced	
	S_b (%)	S_i (%)	S_b (%)	S_i (%)
Reorder top-ranking sentences	**43.46**	**47.40**	20.35	24.27
Reorder documents	34.60	37.85	29.17	31.16
Re-search Web	36.26	38.46	**34.06**	**33.96**

7.2.3. Search strategy reversals

Subjects are given the option to reverse the action in both systems. In Table 5 we give the proportion of each type of action that was undone. We regard this reversal to be an indication of dissatisfaction with the outcome of the action, or in the implicit feedback system's case, the outcome of the action and the terms suggested. Table 5 shows the percentage of reversed actions per search session.

Subjects responded well to the search strategy employed by the implicit feedback system on their behalf. The differences between the subject groups was marked. Inexperienced subjects disliked the effects of the top-ranking sentence reordering, perhaps since they were not used to working with such representations and had different expectations of the action's effect than was actually delivered. Experienced subjects, in contrast liked the top-ranking sentence re-ranking, but reversed the re-searching operation most often. The re-searching occurred in the background, as the subject was viewing representations and exploring the current result set.

7.3. Relevance paths and content

Both systems present a high level of content at the results interface and use relevance paths. In this section we report results on each of these interface features. As there are no path and content differences *between* systems, we only compare results between groups of searchers (i.e., inexperienced vs. experienced).

7.3.1. Relevance paths

Subjects were asked to rate (on a Likert scale) the worth of following a relevance path from one representation of a document to another. The responses were on a scale of 1–5, with a value of 1 reflecting greater agreement (Table 6).

The results show, using MANOVA, that the relevance paths were significantly more *helpful, beneficial, appropriate* and *useful* to experienced subjects than inexperienced. [22] Subjects also felt the distance travelled along the relevance path was a good indicator of the relevance of the information in that path and the document the path came from. Since the Binary Voting Model relied on the viewing of representations to provide it with the evidence it needs to train itself, the success of these interface components was vital to the success of the implicit approach.

Subject interaction with relevance paths was automatically logged. Table 7 shows the most common path taken, the average number of steps followed, the average number of complete and partial paths and the average number of times a subject went straight to a document from the first representation they visited. All averages are for each group of subjects over all tasks. A complete path involved a subject visiting all five document representations and *then* the document itself. Since the most common path was the same for each subject group we do not show the split between systems in the second row of Table 7.

[22] $F_{3,43} = 4.43$, $p = .0084$.

Table 6
Subject perceptions of relevance paths

Semantic differential	Subject group	
	Inexperienced	Experienced
Helpful	2.54	1.97
Beneficial	2.66	1.93
Appropriate	2.34	1.95
Useful	2.62	2.12

Table 7
Use of relevance paths, per task

Factor	Subject group			
	Inexperienced		Experienced	
Most common path	TRS → Title → Summary		TRS → Title → Summary → Summary sentence	
	S_b	S_i	S_b	S_i
Number of steps	3.04	3.17	4.43	4.55
Complete (partial) paths	5.16 (11.54)	5.23 (11.25)	9.43 (17.74)	9.75 (18.15)
Direct to document	5.65	5.39	8.76	8.43

Subjects used relevance paths consistently, although experienced subjects followed the paths for longer. In general, the experienced subjects interacted more with the retrieved documents and more frequently used the document representations as a means of viewing the full-text of a document. Experienced subjects may be more able than inexperienced subjects to adapt to the new interface and the use of document representations and relevance paths. The differences are all significant. [23] There are no significant inter-system differences within each of the subject groups. The nature of the system (i.e., implicit or baseline) does not affect the use of relevance paths.

7.3.2. Content

Both the implicit and manual baseline systems present a large amount of content at the results interface (Figs. 1 and 6). Subjects were asked to express their opinion of this content in the post-task questionnaires and informally at the end of the search session. Inexperienced subjects reacted most positively to this content, as they felt it enabled them to make more accurate relevance assessments. Subjects were also asked whether they felt that showing multiple representations of the same document increased their awareness of document content. The results suggested the content-rich interface was liked by subjects and the increased levels of content facilitated access to more potentially useful information.

8. Discussion

The techniques described in this article gather relevance information unobtrusively from searcher interaction and make decisions on searchers' behalf to reduce the cognitive burden and help them in their seeking.

[23] Number of steps ($t_{22} = 2.34$, $p = .0143$), complete paths ($t_{22} = 3.96$, $p = .0003$), partial paths ($t_{22} = 5.01$, $p < .0001$), direct to document ($t_{22} = 4.53$, $p < .0001$).

Selecting worthwhile terms on behalf of searchers relies on an ability to predict their information needs to a very fine level of granularity. Traditional relevance feedback systems extract terms for query modification from sets of documents (Salton & Buckley, 1990). This approach is coarse-grained and as such is likely to produce a certain number of erroneous terms (i.e., not all terms in a relevant document will actually be relevant). In our approach, we utilise interaction with a novel content-rich interface. The interface uses query-relevant document *representations* to facilitate access to potentially useful information and allow searchers to closely examine results.

From observations and informal post-search interviews, subjects appeared to use the relevance paths and find the increased level of content shown at the result interface of value in their search. This is important, as the success of both systems—especially the implicit version—is dependent on the use of these interface features.

Experienced subjects made more use of the relevance paths. Such subjects may be able to adapt to the new interface technology more easily. However, the content-rich results interface increased inexperienced subject awareness of document content significantly more than experienced subjects. [24] Experienced subjects may be able to infer more from standard representations such as document title and URL and therefore need less information at the interface. Although inexperienced subjects did not use the paths as often (since they were perhaps unfamiliar with the concept), they seemed to prefer the increased levels of content when they did.

The Binary Voting Model chose terms to represent the information needs of the subject. We used the degree of term overlap as a measure of how effectively the model approximated the information needs of subjects. Across all subject groups terms chosen by the Binary Voting Model co-occurred with any subject terms on 72.1% of occasions. On approximately two thirds of the 27.9% occasions that implicit terms *did not* co-occur the decision and background searches, both ill-defined tasks, were being attempted. In these tasks, some subjects expressed a difficulty in generating search terms as they were unfamiliar with the topic. Therefore the system may be selecting terms that the subject had not yet considered yet may be of use in better representing their need. Without prior knowledge of subject search experience and topic knowledge, it was not possible to develop search tasks that were familiar to all subjects.

As subjects did not rate their own search terms as *always* useful, they acknowledge that they are not able to adequately conceptualise their information need, even when given the chance to refine the terms used to express it. However, as they view and process information, and their state of knowledge changes, they become more able to express these needs. The Binary Voting Model, through a process of *reinforcement learning* (Mitchell, 1997) (i.e., being repeatedly shown indications of what constitutes relevance) learns in a similar way, training itself with subject interaction to more fully understand what is relevant.

The form of implicit feedback proposed in this article is at the extreme end of a spectrum of searcher support. Based on informal feedback received during and after the experiment, the approach removed too much searcher control. Implicit feedback systems of this nature may be best used to make decisions in conjunction with, not in place of, the searcher. As in *interactive query expansion* (Koenemann & Belkin, 1996), the system would monitor interaction and present potentially useful terms at the interface. In this collaboration, the searcher, who is best equipped to make relevance decisions, would select potentially useful terms from those offered to modify the query. In our approach it is also conceivable that the system could recommend a retrieval strategy based on the estimated change in information needs. The searcher would have control over whether the recommended action is then executed.

All subjects were instructed before the experiment that the different strategies provided varying degrees of interface support and will have an increasingly dramatic effect on reshaping the information space. They

[24] Participants were asked (on a 5-point Likert scale) whether the content-rich interface improved their awareness of the content of the top-ranked document set (over the presentation in standard web search interfaces). The findings were independent of system used and significant with a paired t-test ($t_{22} = 1.82$, $p = .0412$).

were not told that the control related in any way to shifts, changes or developments in their information need. Subjects adapted well to the need tracking, and seemed comfortable with choosing between the different retrieval strategies.

The approach tracked potential changes or developments in the information need based on changes in the document representations viewed by the searcher. The system communicated its estimations of these developments through the action it performed on the searcher's behalf. Even though the experiment was task-oriented, and used the tasks to engender interaction, the system still estimated shifts or developments in information needs. However, as the results in Fig. 8 show, most of the changes were slight, with only a few medium and large changes detected. We posit that in situations where the need changes dramatically (i.e., from one topic to another unrelated one) there will be increases in the number of document reorders and re-searches. The use of different types of tasks and transitions between tasks make interesting scope for future work.

The baseline system performed better than the implicit feedback system. This was expected, as subjects suggested they would rather have control over the system, and regarded the decisions they made as more accurate and reliable than the implicit feedback systems. This may not always be true (Ruthven, 2003) and perhaps searchers would be wise to place more *trust* in the system. Subjects also suggested that they would like to know *why* the action was taken and *why* certain expansion terms were chosen. The terms may not fit well with the subject's own perception of their need and may still be helpful, or may simply be erroneous. This is an important point, if implicit feedback systems are going to work on behalf of the searcher, it seems reasonable that they also describe the rationale behind their decisions. At present, implicit feedback systems assume a 'black box' approach to assisting those they are meant to help. Explanations may open this box, help engender trust in system actions and perhaps bridge the gap between searcher and system (Ruthven, 2002).

The approach presented in this article has the potential to alleviate some of the problems inherent in explicit relevance feedback whilst preserving many of the benefits that underlie the approach. The initial query is still modified to become more attuned to a searcher's need based on an iterative process of feedback. However, there are three key differences; searchers do not have to explicitly assess and mark documents relevant; these documents are not the finest level of granularity; and the way the new query is used depends on the extent to which the information need has changed (i.e., we do not simply re-search).

The success of the approach bodes well for the construction of effective implicit RF systems that will work in concert with the searcher. To approximate current needs we do not use traditional, potentially unreliable (Kelly & Belkin, 2001), implicit sources of searcher preference (e.g., document reading time, scrolling), but interaction with granular document representations and paths that join them. Unobtrusively monitoring searcher interaction with content-rich interfaces such as that presented in this article may provide a means by which the potential of implicit RF can be realised.

9. Summary

In this article we have presented an implicit feedback approach that uses unobtrusive monitoring of interaction to help searchers find relevant information. To facilitate interaction, we use multiple representations of top-ranked documents, linked by an interactive relevance path. Terms that are chosen from these representations and implicitly weighted using a Binary Voting Model appear to approximate the current information need of searchers. Our approach uses statistical methods to gather evidence and track changes in information needs, tailoring the retrieval strategy to suit the estimated extent of this change.

We conducted a user experiment comparing our approach with a baseline where the searcher was responsible for selecting new query terms and estimating the degree of change in the information need.

The results show that our approach selects terms that are useful and the evidence produced from tracking changes in the information need results in retrieval strategies that were apt and liked by subjects.

Appendix A. T4.Decision

Simulated work task situation: You have recently inherited a large sum of money left by a recently deceased distant relative. A number of friends have advised you that it may be worth investing this money in a financial instrument, such as a bond or corporate stocks. At present you are unaware of stock market trends and lack the knowledge required to make a sound judgement on what to do with this money. You would like information to help you decide.

Task: Bearing in mind this context, your task is to find information that will aid your decision on the *best* type of financial instrument to invest in.

References

Barry, C. L. (1998). Document representations and clues to document relevance. *Journal of the American Society for Information Science, 49*(14), 1293–1303.

Bates, M. (1989). The design of browsing and berry-picking techniques for the online search interface. *Online Review, 13*(5), 407–424.

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., & 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., Oddy, R. N., & Brooks, H. M. (1982). Ask for information retrieval: Part I. Background and theory. *Journal of Documentation, 38*(2), 61–71.

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

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3).

Brajnik, G., Mizzaro, S., & Tasso, C. (1996). Evaluating user interfaces to information retrieval systems: a case study of user support. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 128–136).

Brajnik, G., Mizzaro, S., Tasso, C., & 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.

Bruce, H. W. (1994). A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science, 45*, 142–148.

Callan, J. P. (1994). Passage-level evidence in document retrieval. In *Proceedings of the 17th annual ACM SIGIR conference on research and development in information retrieval* (pp. 302–310).

Campbell, I., & Van Rijsbergen, C. J. (1996). The ostensive model of developing information needs. In *Proceedings of the 3rd international conference on conceptions of library and information science* (pp. 251–268).

Chen, H., & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval system design. *Information Processing and Management, 27*(5), 405–432.

Claypool, M., Le, P., Waseda, M., & Brown, D. (2001). Implicit interest indicators. In *Proceedings of the 6th international conference on intelligent user interfaces* (pp. 33–40).

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

Harman, D. (1988). Towards interactive query expansion. In *Proceedings of the 11th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 321–331).

Hill, W. C., Hollan, J. D., Wroblewski, D., & McCandless, T. (1992). Edit wear and read wear. In *Proceedings of the conference on human factors in computing systems* (pp. 3–9).

Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham.

Jansen, B. J., Spink, A., & 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.

Joachims, T., Freitag, D., & Mitchell, T. (1997). Webwatcher: a tour guide for the world wide web. In *Proceedings of the 16th joint international conference on artificial intelligence* (pp. 770–775).

Kelly, D. (2004). *Understanding implicit feedback and document preference: a naturalistic user study*. Unpublished doctoral dissertation, Rutgers University, New Jersey.

Kelly, D., & Belkin, N. J. (2001). Reading time, scrolling and interaction: Exploring sources of user preferences for relevance feedback during interactive information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 408–409).

Kim, J., Oard, D. W., & Romanik, K. (2000). *Using implicit feedback for user modelling in internet and intranet searching*. College Park: College of Library and Information Services, University of Maryland.

Koenemann, J., & Belkin, N. J. (1996). A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 205–212).

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Reidl, J. (1997). Grouplens: applying collaborative filtering of news. *Communications of the ACM, 40*(3), 77–87.

Kuhlthau, C. C. (1993). *Seeking meaning: a process approach to library and information science*. Norwood, NJ: Ablex Publishing.

Lieberman, H. (1995). Letizia: an agent that assists web browsing. In *Proceedings of the 14th international joint conference on artificial intelligence* (pp. 475–480).

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Morita, M., & Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th annual ACM SIGIR conference on research and development in information retrieval* (pp. 272–281).

Nichols, D. M. (1997). Implicit ratings and filtering. In *Proceedings of the 5th DELOS workshop on filtering and collaborative filtering* (pp. 31–36).

Pirolli, P., & Card, S. (1995). Information foraging in information access environments. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems* (pp. 51–58).

Robins, D. (1997). Shifts of focus in information retrieval interaction. In *Proceedings of the 65th annual meeting of the American society for information science* (pp. 123–134).

Ruthven, I. (2002). On the use of explanations as a mediating device for relevance feedback. In *Proceedings of the 6th European conference on digital libraries* (pp. 338–345).

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th ACM SIGIR conference on research and development in information retrieval* (pp. 213–220).

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text retrieval systems. In *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 49–58).

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

Spink, A., Griesdorf, H., & 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.

Vakkari, P. (2002). Subject knowledge, source of terms, and term selection in query expansion: an analytical study. *Proceedings of the 24th annual european colloquium on information retrieval research* (pp. 110–123).

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

White, R. W., Jose, J. M., & Ruthven, I. (2003a). An approach for implicitly detecting information needs. In *Proceedings of the 12th annual conference on information and knowledge management* (pp. 504–507).

White, R. W., Jose, J. M., & Ruthven, I. (2003b). A granular approach to web search result presentation. In *Proceedings of the 9th IFIP TC13 conference on human computer interaction* (pp. 213–220).

White, R. W., Jose, J. M., & Ruthven, I. (2003c). 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., & Jose, J. M. (2002a). Finding relevant web documents using top ranking sentences: an evaluation of two alternative schemes. In *Proceedings of the 25th annual ACM SIGIR conference on research and development in information retrieval* (pp. 57–64).

White, R. W., Ruthven, I., & Jose, J. M. (2002b). The use of implicit evidence for relevance feedback in web retrieval. In *Proceedings of 24th BCS-IRSG European colloquium on information retrieval research* (pp. 93–109).