Evaluating Implicit Feedback Models Using Searcher Simulations

RYEN W. WHITE
University of Maryland
IAN RUTHVEN
University of Strathclyde
and
JOEMON M. JOSE and C. J. VAN RIJSBERGEN
University of Glasgow

In this article we describe an evaluation of relevance feedback (RF) algorithms using searcher simulations. Since these algorithms select additional terms for query modification based on inferences made from searcher interaction, not on relevance information searchers explicitly provide (as in traditional RF), we refer to them as *implicit feedback models*. We introduce six different models that base their decisions on the interactions of searchers and use different approaches to rank query modification terms. The aim of this article is to determine which of these models should be used to assist searchers in the systems we develop. To evaluate these models we used searcher simulations that afforded us more control over the experimental conditions than experiments with human subjects and allowed complex interaction to be modeled without the need for costly human experimentation. The simulation-based evaluation methodology measures how well the models learn the distribution of terms across relevant documents (i.e., learn what information is relevant) and how well they improve search effectiveness (i.e., create effective search queries). Our findings show that an implicit feedback model based on Jeffrey's rule of conditioning outperformed other models under investigation.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance feedback*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology*

General Terms: Experimentation

Additional Key Words and Phrases: User simulations, evaluation, relevance feedback, implicit feedback

Initial results were presented in White et al. [2004c].
Authors' addresses: R. W. White, Institute for Advanced Computer Studies, University of Maryland, College park, MD 20742; email: ryen@umd.edu; I. Ruthven, Department of Computer and Information Sciences, University of Strathclyde, Glasgow, Scotland G1 1XH; email: ian.ruthven@cis.strath.ac.uk; J. M. Jose and C .J. van Rijsbergen, Department of Computing Science, University of Glasgow, Glasgow, Scotland G12 8RZ; email: {jj,keith}@dcs.gla.ac.uk.

1. INTRODUCTION

Relevance feedback (RF) (cf. Salton and Buckley [1990]) is an iterative technique through which a searcher's query can be automatically improved by the direct provision of relevance information. When using information retrieval (IR) systems that implement RF techniques, searchers typically have to visit retrieved documents, assess their relevance, and convey this information to the system in the form of relevance assessments. However, these processes may intrude on the information-seeking behavior of searchers, forcing them to make decisions about the relevance of search results that they may be unwilling or unable to make [Furner 2002].

Rather than expecting searchers to explicitly mark documents as relevant, *implicit feedback models* can remove the burdens of traditional RF and make inferences about relevance from searcher interaction [Morita and Shinoda 1994; Kelly and Teevan 2003]. Traditional measures of implicit feedback such as document reading time, scrolling, and other similar interactions can be unreliable and context-dependent [Kelly 2004]. However, as we have shown in earlier work, if implemented carefully at the search interface, implicit feedback can be an effective substitute for traditional explicit RF in interactive search environments [White et al. 2002b].

In our research we have developed search interfaces that more actively engage searchers in the examination of search results than traditional styles of result presentation adopted by commercial search systems such as Google[1] and AltaVista[2] [White et al. 2004a, 2004b]. The information shown to searchers in our search interfaces is extracted from top-ranked documents at retrieval time and is characterized by the presence of search terms (i.e., it is query-relevant), and exploring it allows searchers to closely examine search results. Searchers can interact with *document representations* and follow *relevance paths* between these representations, generating evidence for implicit feedback models. Since searchers interact with more information, this provides an increased quantity of evidence for the RF algorithms, and since information is in the form of document representations, not the full-text of documents, their interaction is potentially more focussed, improving the quality of the evidence. Figure 1 provides an example of such an interface that has been shown to be effective with human searchers in previous work [White et al. 2004a].

Documents are represented at the interface by their full text and a variety of smaller, query relevant representations, created at retrieval time. Document representations include the document title and the query-biased summary of the document; a list of *top-ranking sentences* (TRS) extracted from the top 30 documents retrieved, scored in relation to the query; a sentence in the document summary, and each summary sentence in the context it occurs in the document (i.e., with the preceding and following sentence). Each summary sentence and top-ranking sentence is regarded as a representation of the document. These representations allow searchers to more deeply explore the retrieved information and can combine to form an interactive *relevance path* at the search

[1] `http://www.google.com.`
[2] `http://www.av.com.`

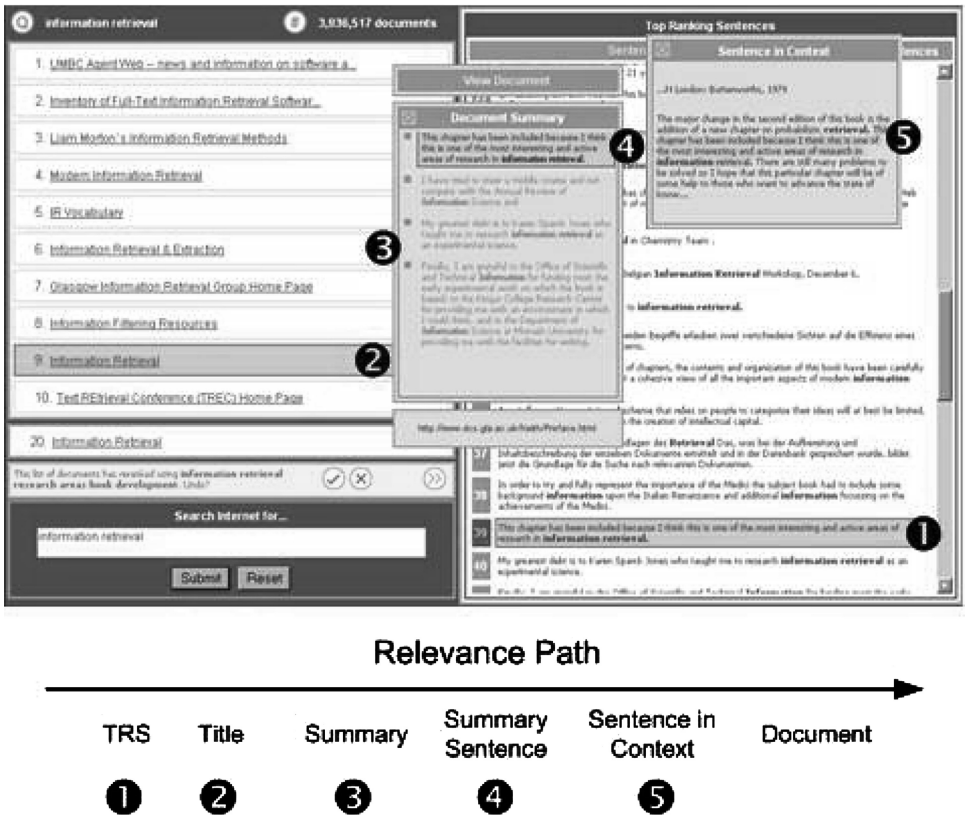


Fig. 1. Example search interface and full relevance path.

interface. The default display in the example interface shown in Figure 1 contains the list of top-ranking sentences and the list of the first 10 document titles.

A relevance path is traversed if searchers travel between different representations of the same document. The paths provide searchers with progressively more information from the best documents to help them choose new query terms and select what new information to view. The presentation of progressively more information from documents to aid relevance assessments has been shown to be effective in related work [Zellweger et al. 2000; Paek et al. 2004]. The further along a path they travel, the more relevant the information in the path is assumed to be. That is, the searcher is implicitly indicating what information in a document is relevant through an examination of the most potentially useful parts. Figure 1 shows a relevance path below the interface.[3] It is through the traversal of these paths that relevance information is communicated to the implicit feedback models.

To evaluate implicit feedback models that operate with this type of interface, we must develop an evaluation methodology that tests their effectiveness when

[3]Numbers below the path correspond to those in the screenshot in Figure 1.

presented with the type of evidence they will be faced with during an interactive search session (i.e., not relevant documents, but relevant sentences, summaries and titles, and paths between them). In this article we describe a study to select the best-performing implicit feedback model from six models that gather relevance information from searcher interaction at the results interface. In White et al. [2004c], we presented the initial results of this evaluation where searcher simulations were used to simulate interaction with interfaces of the type shown in Figure 1. In this article we expand on the description of the work given there and more fully describe the evaluation methodology, test the models in more varied search "situations," present more experimental findings, and discuss the implications of our findings in greater detail. The implicit feedback models and search interfaces we describe are example implementations for experimental purposes and others are possible. The searcher simulations interact with extracted information and provide evidence for each of the implicit feedback models tested in this study; the findings allow us to select the best-performing model.

The remainder of this article describes the process through which the best-performing implicit feedback model was selected. In Section 2 we describe the six implicit feedback models tested in Section 3 the evaluation methodology, and in Section 4 the constraints specific to this study. In Section 5 we present findings on the performance of the implicit feedback models, in Section 6 discuss the findings and their implications, and in Section 7 conclude the article.

2. IMPLICIT FEEDBACK MODELS

In this study we investigated a variety of different methods of RF weighting based on implicit evidence provided through searcher interaction. The implicit feedback models presented use different methods of handling this implicit evidence and updating their understanding of searcher needs in light of it [White, 2004]. The assumption made in the models described in this study is that searchers will try to view information that relates to their needs. That is, they will typically try to maximize the amount of relevant information they view while minimising the amount of irrelevant information [Pirolli and Card 1995]. Simulations provide the models with evidence in the form of representations, relevance paths that join representations, and the full text of documents (i.e., the type of information they will encounter in our search systems). The study compared the models' ability to "learn"[4] relevance and create more effective search queries. We now describe the models in more detail.

2.1 Binary Voting Model

The *Binary Voting Model* [White et al. 2003a] is a heuristic-based implicit feedback model that assumes useful terms will appear in many of the representations that a searcher chooses to view. To identify potentially useful query

[4]The word *learn* is used to refer to the process in which the implicit feedback models improve the quality of their query formulations incrementally during a search session. This process creates a ranking in the list of vocabulary terms that approximates the term distribution across the set of relevant top-ranked documents.

modification terms, the model allows each document representation to "vote" 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. All terms are candidates in the voting process, and these votes accumulate across all viewed representations.

The different types of representation a searcher may view 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 Binary Voting Model. 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, the Binary Voting Model *weights* the contribution of a representation's vote based on the indicative worth of the representations, for example, 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 simply because it is shorter. We used heuristic weights for the indicative worth of each type of representation that ensured that the total score for a term in a relevance path was between 0 and 1 (inclusive).[5]

The terms with the highest overall vote were those taken to best describe the information viewed by the searcher (i.e., those terms that were present most often across all representations) and were used to approximate searcher interests.

2.2 Jeffrey's Conditioning Model

The *Jeffrey's Conditioning Model* [White et al. 2004c] uses Jeffrey's rule of conditioning [Jeffrey 1983] to revise the probability of term relevance in light of evidence gathered from searcher interaction. Jeffrey's conditioning captures the uncertain nature of implicit evidence, and is used since even after the passage of experience (i.e., following a relevance path) the model is still uncertain about the relevance of a term.

In the search interfaces we developed searchers traverse relevance paths between document *representations*, not documents as in other work [Campbell and Van Rijsbergen 1996; Chalmers et al. 1998]. The representations that comprised these paths were smaller than documents, the paths were generally short (i.e., no more than six representations), and the most recent document representation was not necessarily the most relevant. The Jeffrey's Conditioning Model uses a measure of *confidence* to estimate the worth of relevance information; we assigned an exponentially decreasing profile to new relevance information. The assumption made by this model is that the further a searcher travels along a relevance path, the more certain it can be about the relevance of the information towards the start of the path. As the viewing of the *next* representation is exploratory and *driven by curiosity as well*

[5]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.

as information need, the model is cautious, and hence less confident about the value of this evidence.

As well as using the position of a representation in a relevance path as an indication of its value, the quality of evidence in a representation, or its *indicative worth*, can also affect how confident we are in the value of its content. In the Binary Voting Model we used heuristics based on the typical *length* of document representations to measure indicativity. However, titles and top-ranking sentences, which may be indicative of document content, are short and will have low indicativity scores if their typical length is the attribute used to score them. Although the use of representation length is computationally simple, its use may not always be appropriate as a measure of indicativity. Rather than using representation length, the Jeffrey's Conditioning Model uses the set of all nonstopword terms in a representation to compute the indicativity. Each term is assigned a score based on the normalized frequency of its occurrence in the source document. The higher this frequency, the more often a term occurs in the document, and the more representative of document content that term can be seen to be. The summation of the scores for all terms in a representation form an *indicativity index* for that representation.[6]

Confidence measures the worth of a document representation based on its position in the relevance path. *Indicativity* measures the quality of a representation based on how well it represents the concepts from the source document. We computed the *value* of the evidence in a representation by multiplying its indicativity by its confidence. Using the confidence and indicativity measures ensures that the worthwhile representations in each relevance path contribute most to the selection of potentially useful query modification terms. In Equation (1) we show how this measure of value is multiplied by a Bayesian inversion of the standard equation for Jeffrey's Conditioning to compute a revised probability of term relevance. This probability is updated in light of searcher interaction (i.e., the traversal of relevance paths) and after the traversal of a relevance path. The length of a relevance path ranged between one and six steps. We denoted this length using N. When this length is greater than 1, the component updates the probabilities across this path. The probability of relevance of a term across a path of length N is denoted P_N and given through *successive updating*:

$$P_N(t) = \sum_{i=1}^{N-1} c_i . I_i . \left[\left(P_i(t=1|p_i) \frac{P_{i+1}(t=1)}{P_i(t=1)} + P_i(t=0|p_i) \frac{P_{i+1}(t=0)}{P_i(t=0)} \right) . P_i(t) \right], \tag{1}$$

where a representation at step i in the path p is denoted p_i. The confidence in the value of the representation is denoted c_i, and I_i is the indicativity of the representation. This estimation calculates the revised probability of relevance for a term t given a representation p_i, where $P(t=1)$ is the probability of observing t, and $P_i(t=0)$ the probability of not observing t. The prior searcher estimate

[6]This measure is similar to a *Hamming distance* [Hamming 1950], but uses term *weights* rather than presence/absence.

$P_i(t = 1)$ is given by collection statistics (i.e., the normalized term frequency in the top-ranked documents). The probabilities $P_{i+1}(t = 1)$ and $P_{i+1}(t = 1 \mid p_i)$ are computed in the same way as $P_i(t = 1)$, with one difference in each case: rather than using the frequency of term t in the top documents, $P_{i+1}(t = 1)$ uses the frequency of t in the whole relevance path and $P_i(t = 1 \mid p_i)$ uses the frequency of t in the representation p_i. The updated probability $P_N(t)$ reflects the effect of the passage of experience and is similar to that described by Van Rijsbergen [1992].

2.3 *WPQ*-Based Models

The *wpq* method [Robertson 1990] has been shown to produce effective term rankings for query expansion. The equation for *wpq* is shown below, where the typical values r_t = the number of seen relevant documents containing term t, n_t = the number of documents containing t, R = the number of seen relevant documents for query q, and N = the number of documents in the collection:

$$wpq_t = \log \frac{r_t/(R - r_t)}{(n_t - r_t)/(N - n_t - R + r_t)} \cdot \left(\frac{r_t}{R} - \frac{n_t - r_t}{N - R} \right). \tag{2}$$

In the models described in this article, whole documents *and* document representations such as titles, summaries, and top-ranking sentences can be considered relevant. The *wpq* method is based on probabilistic distributions of a term in relevant and nonrelevant documents. As the values of r_t and R change during searcher interaction, the *wpq*-generated term weights also change. For the study described in this article, we developed three variants of the *wpq* approach. In the following sections, these variants are described.

2.3.1 *WPQ Document Model.* This model uses the full text of documents and assumes that all documents presented to the model (i.e., those that are seen) are in some way relevant. The *wpq* formula is applied to each document and the expansion terms chosen from it. In Equation (2) the values of R = the number of seen documents, r_t = the number of seen documents containing term t, N = the number of top-ranked documents, and n_t = the number of top-ranked documents containing the term t. This approach is effectively a traditional explicit RF model and was included in the study to investigate the effects of using the full text of documents for such feedback.

2.3.2 *WPQ Path Model.* In this model, the terms from each complete relevance path are pooled together and ranked based on their *wpq* score. In Equation (2), we use the variable values of R = the number of seen paths, r_t = the number of seen paths containing term t, N = the total number of paths generated from the top 30 retrieved documents, and n_t = the number of generated paths that contain t. Since it uses terms in the *complete path* for query expansion, this model does not use any path weighting or indicativity measures. This model was chosen to investigate combining *wpq* and complete relevance paths for implicit feedback.

2.3.3 *WPQ Ostensive*[7] *Profile Model.* This model considers each representation in the relevance path separately, applying the *wpq* formula and ranking the terms each representation contains. This model adds a temporal dimension to relevance, assigning a within-path *ostensive relevance profile* [Campbell and Van Rijsbergen 1996] that suggests a recently viewed step in the relevance path is more indicative of the current information need than a previously viewed one. This differs from the Jeffrey's Conditioning Model, which assigns a reduced weight to the most recently viewed step in the path. The *wpq* weights are normalized using such a profile. The model treats a relevance path as a series of representations, and uses each representation separately for *wpq*. In this model, the *wpq* formula uses the values R = the number of seen representations, r_t = the number of seen representations containing term t, N = the number of representations in top-ranked documents, and n_t = the number of representations containing the term t. This model uses an ostensive relevance profile to enhance the *WPQ* Path Model presented in the previous section.

2.4 Random Term Selection Model

The random term selection model assigns a random score between 0 and 1 to terms from viewed representations. At the end of each relevance path, the model ranks the terms based on these random scores and uses the top-scoring terms to expand the original query. This model does not use any path weighting or indicativity measures. This model is a baseline and was included to test the degree to which using any reasonable term-weighting approach affected the success of the implicit feedback. Also, since it did not retain any memory of important terms, documents, or document representations, this model was also expected to experience no learning.

So far in this article we have described the implicit feedback models and the type of search interfaces that these models would be deployed on. In the next section we describe the evaluation methodology that we developed to test the implicit feedback models.

3. SIMULATION-BASED EVALUATION METHODOLOGY

RF techniques have traditionally been evaluated in IR without human subjects using measures of search effectiveness (i.e., precision and recall), monitored over a series of feedback iterations [Harper 1980; Robertson 1986]. Our interfaces, such as that shown in Figure 1, facilitate more interaction than the standard "ranked list" form of result presentation. Typically, the only interaction modeled in standard RF experimentation is the provision of relevance feedback through marking relevant documents over a series of feedback iterations [Buckley et al. 1994]. However, while implicit feedback techniques can be relatively simple [Ruthven et al. 2003], they may also use complex interaction metaphors to elicit searcher intentions; while the interaction modeled in RF

[7]The only similarity to the Ostensive Model of Relevance [Campbell 2000] is the exponentially increasing relevance weight applied to document representations at subsequent temporal locations.

experimentation can be useful to assess many RF algorithms, its simplicity may make it inappropriate for situations where the feedback is gathered through a more complex interaction paradigm.

There is no standard way to evaluate term selection models that require complex or copious searcher interaction with results interfaces. Simulation-based methods have been used in previous studies to test query modification techniques [Harman 1988; Magennis and Van Rijsbergen 1998; Ruthven 2003] or to detect shifts in the interests of computer users [Lam et al. 1996; Mostafa et al. 2003]. These methods are worthwhile since they (i) are less time-consuming and costly than experiments with human subjects, (ii) allow the comparison of IR techniques in different retrieval scenarios, and (iii) maintain control over environmental and situational variables. Simulation-based methods have also been used, among other things, to test the usability of Web sites [Chi et al. 2003] and simulate the hyperlink clicks of Web searchers [Chi et al. 2001]. In this article we use simulation-based approaches in a different way from previous studies: to simulate searcher interaction at the results interface and employ such simulations in the evaluation of feedback algorithms for use in search interfaces.

Although simulation-based approaches cannot be used to directly test the interface from a searcher's perspective, they can test the effectiveness of the models that underlie the interfaces in a variety of circumstances that may influence interface design should weaknesses emerge. The creation of a simulation-based approach to evaluate RF algorithms on a particular interface also ensures that designers of the interface think about how searchers could interact with their system.

The simulation-based approach we developed does not model factors such as type of users, search experience, type of information needs, or the domain in which these simulations are used. Our work in this area is initial and formative and we plan to develop simulations that incorporate such factors in future work. While simulations cannot capture the cognitive processes (including the subjective act of human relevance assessment) that can play a large part in the use and evaluation of IR systems [Cosijn and Ingwersen 2000; Borlund 2003], they can allow for a more complete analysis of the techniques and algorithms that underlie these systems prior to their deployment in experimental interfaces. Designers of search interfaces can use this approach as part of the design process, ensuring that only the algorithms with the best overall performance are included in the interfaces they create. In this study a simulation-based evaluation methodology was used to benchmark such models and choose the best-performing model to be deployed in an interactive RF system.

The simulation assumes the role of a searcher, browsing the results of an initial retrieval. The information content of the top-ranked documents in the first retrieved document set constitutes the information space that the searcher must explore. All interaction in this simulation was with this set and a new information space was never generated. This allowed us to evaluate the performance of the model between searcher-defined query iterations, and how

they will generally be expected to perform in operational environments. In the simulation, searchers were modeled using a number of different strategies: (i) assuming the searchers only viewed relevant/nonrelevant information, that is, follow relevance paths from only relevant or only nonrelevant documents; (ii) assuming they viewed all relevant or all nonrelevant information, that is, followed all relevant relevance paths or followed all nonrelevant relevance paths; (iii) assuming they exhibited differing degrees of "wandering" behavior, that is, trying to view relevant information but also viewing different amounts of nonrelevant information.

The interaction simulated related to that afforded by the search interfaces since we were simulating what searchers could do given this interface. However, this does not invalidate the evaluation methodology: if we have a different interface, we have different simulations. However, an important point is that different search interfaces may provide less relevant information or less consistent information to the RF models, directly influencing their performance. In future work we will investigate how performance is affected by changes to the interface through testing the models with other interface designs.

Although we were not conducting a standard TREC-style evaluation, the use of use TREC relevance assessments was still valid for our study as they were assumed to be independent of the interfaces and the systems that led to the documents being assessed. Although we did need to consider the effects of user and task, this study was aimed at evaluating models in a controlled study, so we needed the same assessments for all systems.

The models were tested based on how well they improved search *precision* (the proportion of retrieved documents that are relevant) and "learn" the distribution of terms across the relevant documents. Since searchers typically exhibit limited interaction with search results [Jansen et al. 2000], it is important to ensure that most of the information they interact with is relevant. For this reason, precision was used as a measure of search effectiveness in this study rather than *recall* (the proportion of relevant documents retrieved).

In this section the simulation-based evaluation methodology is introduced. The system, corpus, and topics used are described in Section 3.1. In Section 3.2 the techniques used to extract the relevance paths are described, and in Section 3.3 the different simulated search scenarios that use the relevance paths are described. In Section 3.4 the relevant distributions and correlation coefficients used to evaluate how well the models learn relevance are presented. The evaluation procedure and a description of the study are given in Sections 3.5 and 3.6, respectively.

3.1 System, Corpus, and Topics

The popular SMART search system [Salton 1971] was used in the experiment to index and search the corpus. The test collection used was the *San Jose Mercury News* 1991 document collection (SJMN) taken from the TREC initiative [Harman 1993]. This collection comprises 90,257 documents, with an average 410.7 words per document (including document title) and an average

Table I. Possible Relevance Path Routes

Document representations					
TRS	Title	Summary	Summary Sentence	Sentence in Context	Total
4	1	1	4	1	16
4	1	1	4		16
4	1	1			4
4	1				4
4					4
	1	1	4	1	4
	1	1	4		4
	1	1			1
	1				1

55.6 relevant documents per topic, and has been used successfully in previous experiments of this nature [Ruthven 2003]. The creation of relevance paths requires documents that contain at least four sentences. However, to create worthwhile paths with well-formed "sentences in context," the component requires documents that contain around 10 sentences.[8]

TREC topics 101–150 were used and the query was taken from the short *title* field of the TREC topic description. The use of the title was appropriate because it was similar in length and content to real user queries. The simulation retrieved the top 30 results for each of the 50 TREC topics used as queries in this study; these results can contain both relevant and nonrelevant documents. In some scenarios, the simulation required paths from only nonrelevant documents, only relevant documents, or a mixture of both. However, for some topics, there were no relevant documents in the top 30 results, making the execution of scenarios that used relevant documents problematic. Therefore, the number of search topics used depended on the scenario.[9] We now explain how paths were extracted from top-ranked results for each topic.

3.2 Relevance Paths

In the simulation, paths were extracted just from relevant documents or from a mixture of relevant and nonrelevant documents, depending on the simulation strategy. Each document had a preset number of representations and number of possible relevance path routes between these representations. In Table I all routes for all path types are shown. The final "document" step was not included in the simulation since it was not used as evidence by the implicit feedback models.

For example, for viewing all five representations (first row of Table I) there were $4 \times 1 \times 1 \times 4 \times 1 = 16$ possible paths.[10] The final column shows the

[8]Documents with only four sentences may result in poor summaries and sentences in context comprised of other summary sentences, not new sentences that may contain useful alternate terms.
[9]For scenarios demanding relevant documents, we used 43 of the 50 topics, and for those demanding nonrelevant documents, we used all 50 topics.
[10]The list of TRS comprised all sentences from the top 30 document *summaries*; these summaries were four sentences long. This length was shown to be effective with real users in earlier work [Tombros and Sanderson 1998] and explains why there are four possible starting points if a relevance path starts from a TRS.

total for each possible route. There were 54 possible relevance paths for each document. If all top 30 documents were used, there were 1,620 (54×30) possible relevance paths. The path ordering was not directed by the user interface, but from the way users employed paths based on previous iterative design. In the next section, more details are given on how search scenarios that used these paths were deployed in the simulation.

3.3 Simulated Search Scenarios

To operate effectively, the implicit feedback models should handle different retrieval situations. Since the models rely on the interaction of searchers, it is necessary to test them with different styles of interaction or retrieval scenarios. To do this, the way in which relevance paths are chosen is varied and the models are tested in *extreme* and *premodeled* situations. In this section, styles of interaction that represent each of these situation categories are described in more detail.

3.3.1 *Extreme Situations.* Styles of interaction in this category represent extreme situations where *only* relevant or nonrelevant paths are traversed. Two strategies are presented, one where all paths are traversed and another where a subset of these paths is traversed. These strategies create bounds on the performance of the system and model the situation where searchers (by chance) interact *only* with relevant or nonrelevant information. They determine the best or worst expected performance of the models, depending on the paths or documents chosen.

—*All paths.* This strategy creates relevance paths from all documents in the top 30 retrieved by the search system. Each relevance path is treated in isolation and the effect of paths traversed in sequence is not cumulative. Although queries submitted for different TREC topics retrieve different numbers of relevant and nonrelevant top-ranked documents, this approach allows the best- and worst-performing paths (and sets of paths) for each topic, and across all topics, to be identified. This approach can be useful to establish the attributes of good and bad relevance paths.

—*Subset of paths.* Searchers would typically not view all retrieved information. This strategy randomly selects a subset of paths used in the "all paths" strategy. Paths are traversed in sequence and the effect across paths is cumulative. That is, unlike the "all paths" strategy, the term scores in the term selection models are not reset after each path. This situation models circumstances where searchers view a number of relevance paths in sequence and all paths viewed contribute in some way toward the weighting of terms for query expansion. Models that perform well in cumulative situations may perform better in real-world RF scenarios where feedback is applied throughout a search session and results marked as relevant are related in some way.

3.3.2 *Premodeled Situations.* The implicit feedback models assume searchers will try to interact with relevant information, but accept that they will

inevitably also view information that is nonrelevant. This assumption is based on our intuition about how searchers generally interact with search systems (i.e., they try to maximize their exposure to relevant information but inevitably view nonrelevant information), the shortcomings of the retrieval algorithms that underlie these systems (i.e., they retrieve nonrelevant information), and the shortcomings of searchers in formulating queries to retrieve this information. Premodeled situations model circumstances where searchers may view relevant and nonrelevant paths as they explore the retrieved information. This level of "wandering" is measured as a percentage of the viewed paths that are not from relevant documents. For the purposes of this study, these paths were regarded as nonrelevant. The effectiveness of the term selection models at different levels of wandering can be tested. The amount of wandering can vary due to search experience or familiarity with the task and the topic of the search. It is possible to vary how relevant (R) and nonrelevant (N) paths are distributed to test how the models perform in different circumstances.

The first path to be visited is chosen at random from the list of available paths. This path can be relevant or nonrelevant. Subsequent paths are randomized in such a way that for 10 paths and 50% wandering the order of traversal may be $\{R, N, R, N, N, R, R, N, R, N\}$. The paths are traversed from the first path onward. The method decides whether the path will be relevant or irrelevant using the order of traversal and selects the actual path based on candidate path quality and its similarity to the current path. The *quality* of a relevance path is measured by its indicativity index used in the Jeffrey's Conditioning Model. The index is a measure of how well a document representation represents the concepts in its source document. The degree to which subsequent paths are *related* is computed using the *Pearson product moment correlation coefficient*. This coefficient has been shown to be an effective measure of similarity in a related study with human subjects [White and Jose 2004; White 2004]. The product of these two measures is used as a decision metric to rank candidate relevance paths and select future paths. The highest-ranked candidate path is chosen as the next path to be traversed. The use of this combined measure simulates searchers' desire to view high-quality, related information. That is, the path with the highest aggregate quality and similarity to the current path is the most likely to be traversed next by a simulated searcher. Relevance is complex and has many possible conceptions [Saracevic 1975; Spink et al. 1998]. Since we are unable to adequately model such conceptions in our simulation, we used the information we do have available (i.e., quality and relatedness) to approximate some aspects of the decision-making process searchers engage in prior to following each relevance path. Figure 2 illustrates this process across the selection of four relevance paths. At successive temporal locations, the searcher must make decisions about what information to view next (shown by "?" in Figure 2). The boundaries between paths appear seamless to the searcher, but are used by the implicit feedback models in deciding when to revise term weights.

In $\{R, N, R, N, N, R, R, N, R, N\}$, the path at position 2 is nonrelevant. To select the second path, all candidate nonrelevant paths are ranked based on the

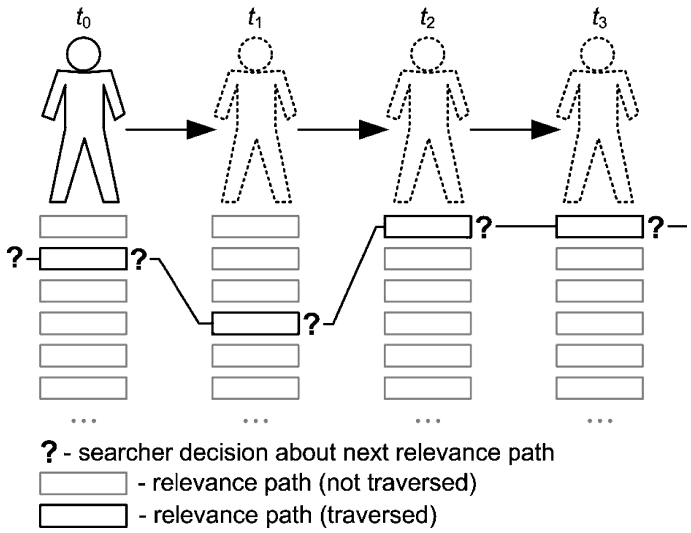


Fig. 2. Selection of relevance paths at successive temporal locations ($t_0 - t_3$).

product of their quality and similarity to the path at position 1. The highest-ranked path is chosen as the next step, and the process repeats until 10 paths have been visited in the order described. Premodeled situations are potentially more realistic than extreme situations since they make real-time predictions on what paths to follow and do not assume that searchers only interact with relevant or nonrelevant information.

3.3.3 *Path Length Distribution.* The modeled situations use empirical evidence to decide that relevance paths taken from irrelevant documents were short, that is, three steps or fewer. However, it is possible to further analyze these results and derive another strategy that creates a distribution of path lengths across relevant and nonrelevant paths. Data gathered from interactive experimentation using a search interface similar to that shown in Figure 1 allowed the construction of path length distributions [White 2004]. A system in that experiment allowed subjects to explicitly mark document representations as relevant during the course of a search. In that experiment, relevance paths were considered as relevant if one or more of their constituent representations were marked as relevant by experimental subjects. Table II shows how path lengths were distributed across relevant and nonrelevant relevance paths.

From these results, it appears that searchers interacted differently with relevant and irrelevant information. More specifically, the results demonstrate that paths were longer if they contained relevant information. The values in Table II can be used in premodeled situations to control the number of paths of each length used in the simulation. For example, if there are 10 relevant paths and 0% wandering that is, $\{R, R, R, R, R, R, R, R, R, R\}$, then there would be one path of length 1 (14.18% of 10), one path of length 2 (9.53% of 10), two of

Table II. Path Length Distribution in Relevant and Nonrelevant Paths (Values Are Percentages)

Steps	Path Type	
	Relevant	Nonrelevant
1	14.18	23.45
2	9.53	25.76
3	18.95	30.28
4	25.11	13.67
5	32.23	6.84

length 3 (18.95% of 10), three of length 4 (25.11% of 10), and three of length 5 (32.23% of 10). The number of paths of each length are rounded to the nearest integer. These path length distributions may be used to simulate the general behavior of real searchers when using content-rich interfaces. This can be a robust alternative to choosing paths regardless of length or imposing upper bounds on the lengths of paths from irrelevant documents.

In all scenarios, model performance is measured based on how the modified queries the models generate influence search precision. As well as being able to improve search effectiveness (through creating well-formed queries), the models should learn relevance when shown examples of what is relevant. In the next section we describe the use of relevant distributions and correlation coefficients to measure such learning.

3.4 Relevant Distributions and Correlation Coefficients

A good implicit feedback model should, given evidence from relevant documents, learn the distribution of terms across the relevant document set. The model should train itself, and become attuned to searcher needs in the fewest possible iterations.

A relevant term space for each topic is created before any experiments are run. This space contains terms from all the relevant documents for that topic, ordered based on their probability of relevance for that topic. After each iteration, the extent to which the term lists generated by the implicit model correlates with the relevant distribution is measured. The simulation "views" relevance paths from relevant documents and provides the models with the implicit relevance information they need to train themselves. We measure how well the models *learn* relevance based on how closely the term ordering they provide matches the term ordering in the relevant distribution.

To measure this, we use two nonparametric correlation coefficients, *Spearman's rho* and *Kendall's tau-b*. These have equivalent underlying assumptions and statistical power, and both return a coefficient in the range $[-1, 1]$. However, they have different interpretations; the Spearman accounts for the proportion of variability between *ranks* in the two lists, the Kendall represents the difference between the probability that the lists are in the same order versus the probability that the lists are in different orders [Siegel and Castellan 1988]. Both correlation coefficients are used to verify learning trends.

3.5 Evaluation Procedure

The simulation creates a set of relevance paths for all relevant and nonrelevant documents in the top-ranked documents retrieved for each topic. The use of these paths, how feedback iterations are generated, and the number of feedback iterations (m) depend on the scenario employed. After each iteration, we monitor the effect on search effectiveness and how closely the ranked list of all possible query modification terms generated by the model correlates with the term distribution across that topic's relevant documents. The correlation is a measure of how well the model learns the relevant term distribution, and precision is a measure of search effectiveness.

The following procedure is used *for each topic with each model*:

(1) use SMART to retrieve a document set in response to a query (i.e., topic title) using an *idf* weighting scheme and record the initial precision values;
(2) identify relevant or nonrelevant documents in the top 30 retrieved documents, depending on the experimental run and store in set s;
(3) select top-ranking sentences from all documents in s using the approach presented in earlier work [White et al. 2003b];
(4) create and store all potential relevance paths for each document in s (up to a maximum of 54 per document);
(5) choose relevance paths or documents as suggested by the simulation strategy, setting m to the number chosen; the Java[11] random number generator is used where appropriate in selecting random paths or documents;
(6) for *each* of the m relevance paths/documents:
    (a) weight terms in path/document with chosen model,
    (b) monitor Kendall and Spearman by comparing order of terms with order in that relevant distribution for that topic,
    (c) choose top-ranked terms and use them to expand original query,
    (d) use new query to retrieve new set of documents, and
    (e) compute new precision values.

To represent a searcher exploring the information space, all simulated interaction was with the results of the first retrieval only. All subsequent retrievals were to test the effectiveness of the new queries and were not used to generate relevance paths.

In this section we have described the methodology we developed to test the implicit feedback models. In the next section we describe the study that uses this methodology.

4. SIMULATION-BASED STUDY

A study of how well each term selection model learned relevance and generated queries that enhanced search effectiveness is now presented. The models were tested in extreme and premodeled situations, and each required a different evaluation approach. The strategies used either the 43 "useable" topics (only

[11] `http://java.sun.com`.

paths from relevant documents) or all 50 topics (only paths from nonrelevant documents) and added six terms to the original query. This was done without any prior knowledge of the effectiveness of adding this number of terms to queries for this collection. Harman [1988] showed that six terms were a reasonable number of additional terms for use in simulated experiments. Query expansion was used to test the marginal effectiveness of the model, that is, how much each new query improved the retrieval over the query before any modification. A *run* in the study involves the testing of a model, under a particular experimental condition. An *iteration* is a single relevance path or document.

4.1 Extreme Situations

The evaluation strategy used in extreme situations modeled the situation where searchers have (by chance) interacted with relevant or irrelevant information.

4.1.1 *All Paths.* This strategy used all paths from the top 30 relevant documents and all paths from the top 30 nonrelevant documents. A run of the simulation comprised $54n$ relevance paths, where n was the number of relevant/nonrelevant documents. The correlation coefficients and search effectiveness were measured after each iteration. The effects of term scoring across consecutive paths was not cumulative. That is, paths were treated in isolation. The evaluation investigated performance differences of paths generated (e.g., best path/worst path, best/worst sets of paths, performance of an average choice of paths).

4.1.2 *Subset of Paths.* This strategy used a subset of the paths generated in the "All Paths" strategy. We ran the simulation 10 times, and *each run comprised 20 iterations*. We recorded correlation coefficients and measures of search effectiveness at iterations 1, 2, 5, 10, and 20. This allowed us to monitor model performance at different points in the search. In the document-centric approach, each *document* is an iteration. Therefore, when this approach was used, it was only possible to have as many iterations as there were relevant/nonrelevant top-ranked documents.

4.2 Premodeled Situations

Three premodeled methods were tested in this study. Unlike the extreme situations, these methods did not assume that searchers could only interact with relevant information. The "Related Paths" method made decisions on what paths to visit based on those traversed previously. In a similar way to the "Subset of Paths" strategy, we ran the simulation 10 times for each implicit model and recorded correlation coefficients and measures of search effectiveness at iterations 1, 2, 5, 10, and 20. The level of wandering was varied in each of the models and recorded at 10%, 20%, 30%, 40%, and 50%. In the document-centric approach, the minimum amount of wandering was one document. Across all premodeled situations, the effect of path length could be ignored or path length distributions based on the results of empirical studies used to make more informed path choices.

Table III. Experimental Scenarios and Variation in Experimental Variables

Scenario		Paths/documents		Relevance			Path Length Distribution	Wandering
Number	Name	All	Subset	R	N	R and N		
1	All Paths	•		•				
2	All Paths	•			•			
3a	Subset of Paths		•	•				
3b	Subset of Paths		•	•			•	
4a	Subset of Paths		•		•			
4b	Subset of Paths		•		•		•	
5a	Related Paths		•			•		•
5b	Related Paths		•			•	•	•

4.3 Experimental Scenarios

In this section we describe the eight experimental scenarios that tested the implicit feedback models in different circumstances. Table III shows these scenarios and the variables changed in each scenario. If a variable varied as part of a scenario, a dot (•) is shown in the corresponding cell.

Scenarios 3, 4, and 5 were each divided into scenarios "a" and "b". In "a" paths were selected randomly, whereas in "b" a path length distribution was used to select paths. In each scenario, all six implicit feedback models introduced earlier in this article were used to generate new queries. The resultant precision values and correlation coefficients were used to assess the performance of the models.

In the next section we describe the results of the simulated study for each experimental scenario with each implicit feedback model.

5. RESULTS

The study was conducted to evaluate a variety of implicit feedback models using searcher simulations. In this section we present the results of our study for each simulated scenario. In particular, we focus on results concerning search effectiveness and relevance learning. We use the terms *bvm*, *jeff*, *wpq.doc*, *wpq.path*, *wpq.ost*, and *ran* to refer the Binary Voting, Jeffrey's Conditioning, *wpq* document, *wpq* path, *wpq* ostensive, and random models, respectively.

In this section we use a number of metrics to assess the performance of the implicit feedback models. In Table IV we identify these metrics, how their values were interpreted, and what values reflected a positive result.

In the remainder of this section we present results of scenarios that used these metrics.

5.1 Scenario 1: All Relevant Paths

The aim of this scenario was to predict the best- and worst-performing paths for each model. In this scenario, all extracted paths across all relevant documents for each topic were used on a per topic basis. For each topic, there were $54n$ paths, where n was the total number of relevant documents in the top 30 retrieved. In total, there were 15,174 paths (i.e., 54×281[12]) across the 43 topics

[12]In total, there were 281 relevant documents in the top 30 retrieved for all 43 search topics used.

Table IV. Metrics Used to Assess Models and Their Interpretation

Measure	Interpretation	Positive Result
Document indicativity	How representative a relevance path is of its source document.	High
Distribution indicativity	How representative a relevance path is of the relevant term distribution.	High
Precision	How an implicit feedback model affects search effectiveness.	High
Marginal precision	How an implicit feedback model affects search effectiveness on each recorded query iteration.	High
Correlation	How an implicit feedback model learns what information is relevant.	High
Marginal correlation	How a relevance path affects the rate in which the implicit feedback model learns the relevant term distribution.	High
Standard deviation	How robust an implicit feedback model is across queries on different search topics.	Low

Table V. Average Best Path Performance in Scenario 1

Term Selection Model	Rank Order	Marginal Correlation	Length	Number of Terms	Indicativity	
					Document	Distribution
bvm	4	0.580	3.9	186 (45.6%)	0.391	0.076
jeff	1	0.659	3.1	139 (47.0%)	0.448	0.062
wpq.doc	3	0.616	—	—	1.000	0.049
wpq.path	2	0.640	3.9	146 (46.9%)	0.632	0.045
wpq.ost	5	0.529	3.9	158 (45.3%)	0.517	0.049
ran	6	0.503	4.0	172 (47.7%)	0.364	0.062

Table VI. Average Worst Path Performance in Scenario 1

Term Selection Model	Rank Order	Marginal Correlation	Length	Number of Terms	Indicativity	
					Document	Distribution
bvm	4	−0.278	3.5	141 (48.8%)	0.295	0.045
jeff	1	−0.219	3.5	168 (44.7%)	0.366	0.043
wpq.doc	6	−0.594	—	—	1.000	0.033
wpq.path	5	−0.289	4.3	179 (47.7%)	0.386	0.030
wpq.ost	2	−0.253	3.1	130 (45.9%)	0.411	0.053
ran	3	−0.264	4.3	172 (46.7%)	0.323	0.040

used in this study. After each path, the effect of that path on correlation coefficients was recorded, and for each model the 15,174 paths were ranked based on their marginal effect on the Spearman and Kendall correlation coefficients. That is, the paths were ranked independent of source document, based on their ability to increase the rate in which the term selection model learned relevance. This allowed us to predict the 10 best- and worst-performing paths and analyze why some paths were good and some were bad. In Tables V and VI, we show the average best and worst path performance for each of the six term selection models, including the marginal effect on correlation (averaged across both

coefficients) of each path, the average path length, and the indicativity score in relation to the source document and the relevant distribution the model is trying to learn. In Table V we also show total number of terms in a path and in brackets the percentage of those terms that are stop words (i.e., common words such as *a*, *the*, *of*).

The same paths perform differently for different term selection models, and only very rarely does the same path appear as the best path for a number of models. The ability of a term selection model to learn what information is relevant is dependent on the paths used. A good term selection model should maximize the rate of learning when shown relevant information, but minimize the negative effects when shown irrelevant information.

Path length, the number of terms, and the percentage of those terms that were stop words had little influence over path performance. However, the indicativity, or *quality*, appeared different between good- and bad-performing paths. We can conjecture from this that paths that lead to poor term selection model performance are not indicative of their source documents or the relevant term distribution for the TREC topic they were created relative to. These results also describe the best and worst possible correlation values for each of these models. The Jeffrey's Conditioning and wpq.path models performed best, as they had the highest potential marginal gains in correlation coefficients and the lowest potential marginal losses for selecting random path from the set of all paths.

5.2 Scenario 2: All Nonrelevant Paths

This scenario was very similar to Scenario 1 but used paths from nonrelevant rather than relevant documents. This was meant to model the situation where, by chance, searchers had viewed all paths from nonrelevant documents. We used the top-ranked sentences from the nonrelevant documents to create the representations that comprised the relevance path. We used these sentences as nonrelevant information and not, say, the bottom-ranked sentences from nonrelevant documents. This is potentially more realistic, as when used in real retrieval situations a search system implementing these techniques will always use top-ranked sentences to form document representations, regardless of whether the documents are relevant or nonrelevant.

In total, there were 65,826 possible path routes (i.e., 54×1219[13]) for each of the six term selection models tested. The paths were again ranked based on the marginal correlation coefficient effects and the best- and worst-performing 10 paths chosen for this analysis. As suggested earlier in this article, the paths chosen from negative documents were assumed to be shorter than relevant paths. For each model, Tables VII and VIII show the average path performance, the average number of terms, and the proportion that were stopwords.

[13]In total, there were 1219 nonrelevant documents in the top 30 retrieved for all 50 search topics used.

Table VII. Average Best Path Performance in Scenario 2

Term Selection Model	Rank Order	Marginal Correlation	Length	Number of Terms	Indicativity	
					Document	Distribution
bvm	4	0.303	3.9	144 (45.5%)	0.258	0.010
jeff	2	0.392	3.5	165 (44.7%)	0.507	0.029
wpq.doc	1	0.434	—	—	1.000	0.025
wpq.path	6	0.239	3.7	146 (47.0%)	0.294	0.008
wpq.ost	3	0.332	3.4	139 (47.7%)	0.220	0.007
ran	5	0.244	4.0	163 (47.1%)	0.176	0.013

Table VIII. Average Worst Path Performance in Scenario 2

Term Selection Model	Rank Order	Marginal Correlation	Length	Number of Terms	Indicativity	
					Document	Distribution
bvm	3	−0.478	3.6	150 (46.6%)	0.203	0.010
jeff	2	−0.433	3.8	168 (46.8%)	0.388	0.027
wpq.doc	6	−0.627	—	—	1.000	0.024
wpq.path	5	−0.517	3.5	142 (42.7%)	0.246	0.004
wpq.ost	1	−0.416	3.7	160 (46.3%)	0.254	0.005
ran	4	−0.513	3.9	147 (50.3%)	0.188	0.008

The Jeffrey's Conditioning and wpq.doc models outperformed the other term selection models. However, the wpq.doc model appeared most variable, with the highest marginal gains but also the highest losses. In a similar way to Scenario 1, the indicativity of the relevant document distribution is a good measure of the quality of the relevance path. Also, since the paths were taken from nonrelevant documents, the indicativity of the relevant distribution (created from relevant documents) was lower than paths from relevant documents, shown in Tables VII and VIII. Also, for paths from nonrelevant documents, there appeared to be no association between path performance and relevant distribution indicativity.

For Scenario 1 and Scenario 2, we did not measure precision after each path. Across relevant and nonrelevant documents, there were approximately 81,000 paths in total. It was not feasible to run all paths through the SMART system to determine marginal precision effects. In Scenarios 3a–5b, we demonstrate a close relationship between the rate of learning and measures of precision. In situations where it may not be practical to compute precision, the correlation coefficients may be a reasonable approximation. In Scenario 2 (as in Scenario 1), the path length, the number of terms, and the number of those terms that were stopwords appeared to have no effect on path performance.

5.3 Scenarios 3a and 3b: Subset of Paths

The relevant subset strategy used a set of relevance paths taken from the top-ranked relevant documents. This scenario models the situation that may arise out of chance if all the information a searcher views is from documents that were relevant.

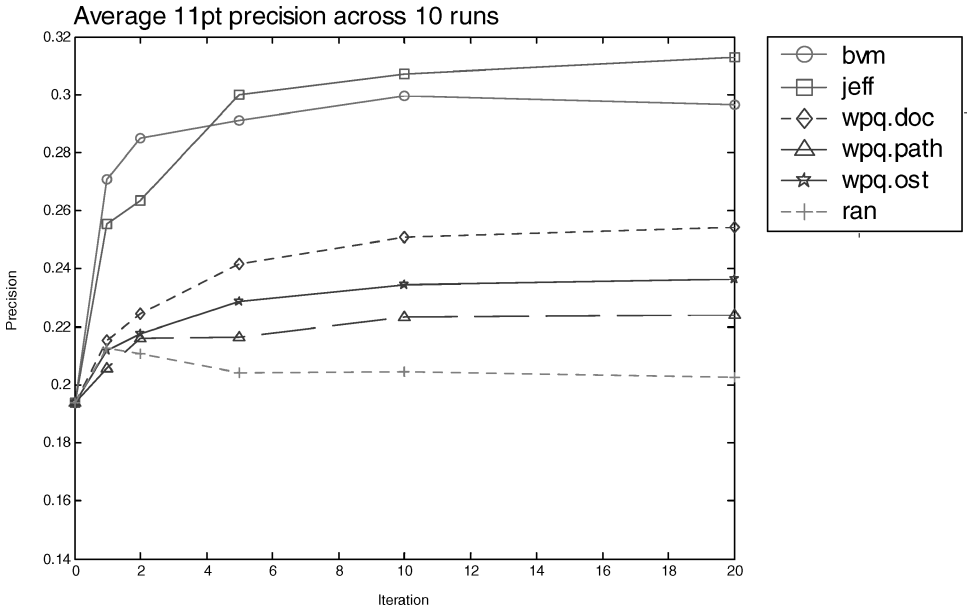


Fig. 3. Average 11-point precision across 10 experimental runs in Scenario 3a.

5.3.1 *Search Effectiveness.* In Scenario 3a, we measured search effectiveness for each of our implicit models through their effects on precision. Figure 3 shows the average 11-point[14] precision values for each model across all iterations. As the figure illustrates, all models increased precision as the number of iterations increases.

Figure 3 presents the actual precision values across all 20 iterations. The Jeffrey's Conditioning and Binary Voting Models outperformed the other implicit feedback models, with large increases inside the first five iterations. Both models were quick to respond to implicit relevance information, with the largest marginal increases (change from one iteration to the next) coming in the first iteration. The other models did not perform as well, but steadily increased until around 10 iterations, where precision leveled out.

Table IX illustrates the marginal difference more clearly than Figure 3, showing the percentage change overall and the marginal percentage change at each iteration.

As Table IX shows, the largest increases in precision came from the Binary Voting Model and the Jeffrey's Conditioning Model, although after 20 iterations the marginal effects of all models appeared slight. The random model performed poorly, although still leading to small overall increases in precision over the baseline. Even though the *random model* assigned each term a random score, the paths selected by the simulation were still query-relevant. Our results show that choosing terms randomly from relevance paths can help improve short queries to a small degree.

[14]The average precision across 11 *recall* values ranging from 0.0 to 1.0, with an increment of 0.1.

Table IX. Percentage Change in Precision per Iteration in Scenario 3a (Overall change in first column, marginal change in second shaded column. Highest percentage in each column in bold.)

	Iterations									
Model	1		2		5		10		20	
bvm	**28.4**	—	**31.9**	**+4.9**	33.4	+2.9	35.3	+2.9	34.6	−1.1
jeff	24.1	—	26.4	+3.0	**35.3**	**+12.2**	**36.9**	+2.4	**38**	**+1.8**
wpq.doc	10	—	13.6	+4.1	19.8	+7.1	22.8	**+3.7**	23.7	+1.2
wpq.path	5.8	—	10.2	+4.6	10.4	+0.2	13.2	+3.2	13.4	+0.2
wpq.ost	8.5	—	10.9	+2.6	17.2	+4.8	17.2	+2.5	18	+0.9
ran	8.8	—	7.9	−1.1	5.0e	−3.1	5.3	+0.2	4.2	−1.1

The *wpq*-based models appeared to follow a similar trend. At each iteration a one-way repeated-measures analysis of variance (ANOVA) was carried out to compare all three *wpq*-based models and t-tests for pair-wise comparisons where appropriate. During the first two iterations, there were no significant differences (iteration 1: $F(2, 27) = 2.258$, $p = .12$; iteration 2: $F(2, 27) = 1.803$, $p = .18$) between the *wpq* models tested. ANOVAs across iterations 5, 10, and 20 suggested there were significant differences in precision between the three *wpq*-models. A series of t-tests revealed the *WPQ* Document Model performed significantly better than both path-based *wpq* models (ostensive-path and path) for iterations 5, 10, and 20 ($p < 0.05$). The relevance paths were not of sufficient size and did not contain a sufficient mixture of terms from which *wpq* could choose candidates for query expansion.

5.3.2 *Relevance Learning.* How well the implicit models trained themselves when given relevance information by the simulation was measured. This was done through the degree of correlation between the ordered list of terms in the topic's relevant distribution and the ordered list of terms chosen by the implicit model. Figure 4 shows the average Spearman and Kendall correlation coefficients across all 43 topics.

Both coefficients followed similar trends for all implicit feedback models. Again, the Jeffrey's Conditioning Model and Binary Voting Model learned at a faster rate, with the model based on Jeffrey's rule of conditioning performing best. The random model returned a coefficient value close to zero with both coefficients. In both cases, a value of zero implies no correlation between the two lists, and this was to be expected if the model randomly ordered the term list. For all other models, the coefficients tended to 1, implying that the models were *learning* the relevant distribution from the given relevance information. Both the Jeffrey's Conditioning Model and the Binary Voting Model obtained high levels of correlation after the first iteration, whereas the *wpq* models needed more *training* to reach a level where the terms they recommended appeared to match those in the relevant distribution.

In Scenario 3b, the paths were chosen at random from the set of paths extracted from relevant documents. However, the path length distribution was used to control the number of paths of different lengths that were used in the simulation. The resulting findings of this scenario demonstrated little difference with the random paths approach used in Scenario 3a.

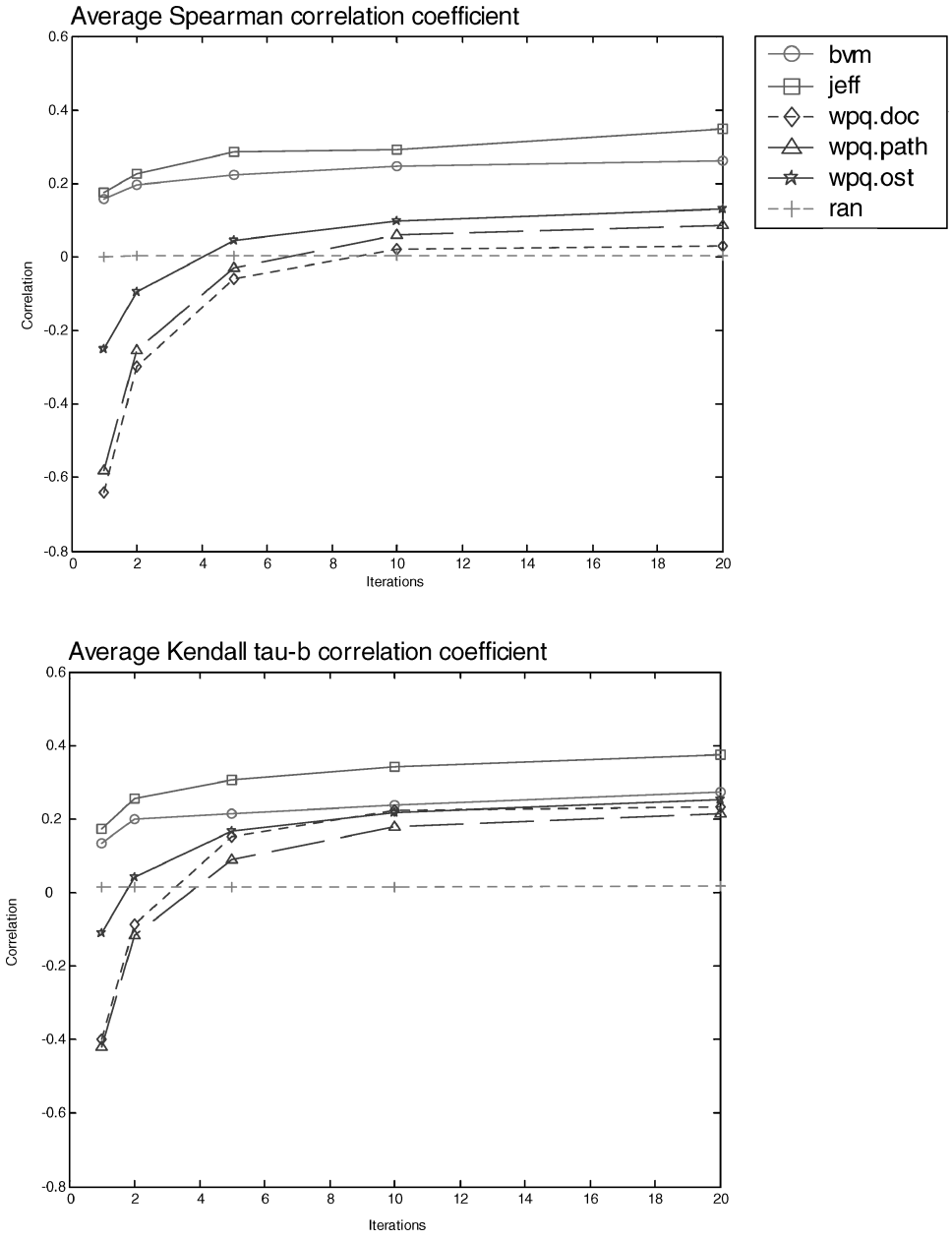


Fig. 4. Average correlation coefficient values across 10 experimental runs in Scenario 3a.

5.4 Scenarios 4a and 4b: Subset of Paths

Scenarios 4a and 4b, in a similar way to Scenarios 3a and 3b, used a subset of available paths. This scenario modeled the situation that may arise if, by chance, all the information a searcher views is from documents that are non-relevant. It is reasonable to assume that searchers will view *some* information

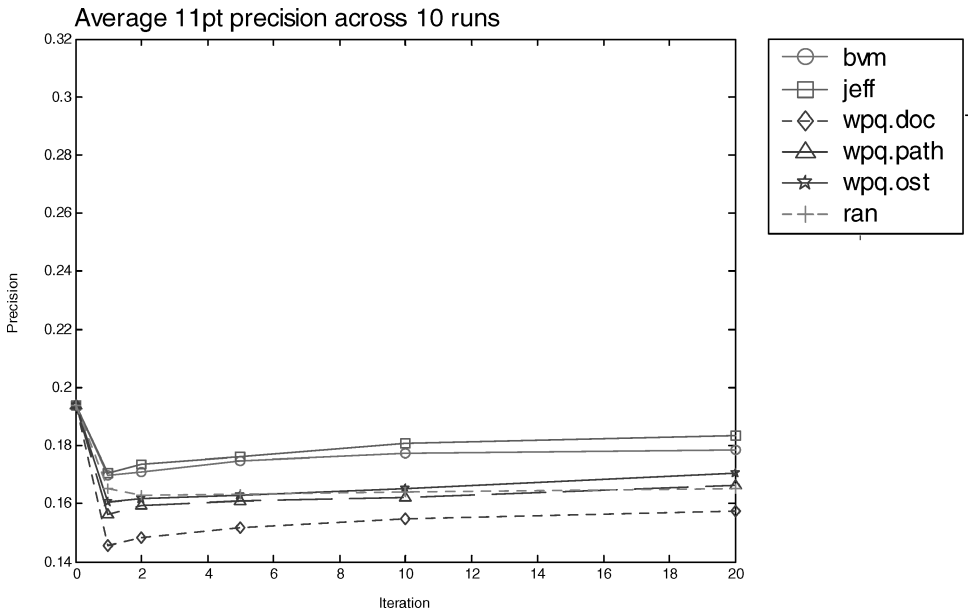


Fig. 5. Average 11-point precision across 10 experimental runs in Scenario 4a.

from nonrelevant documents as they search. It is only in extreme situations where *all* the information they view is from nonrelevant documents. These scenarios model such an extreme situation.

5.4.1 *Search Effectiveness.* We measured search effectiveness for each of our implicit models through their effects on precision. Figure 5 shows the average 11-point precision values for each model across all 20 iterations. All models increased the precision after the first iteration; however, as the figure illustrates, some models increased overall precision and some reduced overall precision.

The Jeffrey's Conditioning and Binary Voting Models outperformed the other implicit feedback models. Although the increases in precision were small, the Jeffrey's Conditioning and Binary Voting Models seemed better able to create effective search queries in situations where relevant information was difficult to find. That is, they seemed better able to use paths from nonrelevant documents to select terms for query modification. The other models did not perform as well, but steadily increased until around 10 iterations, where precision leveled out.

The paths from nonrelevant documents typically contain very few or no query terms. The relevance paths are sentence-based and sentences are scored based on the algorithm for scoring top-ranking sentences described in White et al. [2003b]. A large proportion of each sentence's score is derived from its relation to the query. If there are few query terms, then other factors, such as the location of a sentence in a document and any words in the document that also appear in the document title are used to weight relevance paths. The paths chosen are therefore document-dependent, not query-dependent, and may cover a number of unrelated themes. While all the models appeared to be affected

Table X. Percentage Change in Precision per Iteration in Scenario 4a (Overall change in first column, marginal change in second shaded column. Highest percentage in each column in bold.)

Model	Iterations									
	1		2		5		10		20	
bvm	−14.4	—	−13.5	+0.8	−11.1	+2.1	−9.2	+1.7	−8.6	+0.6
jeff	**−13.7**	—	**−11.8**	+1.8	**−10.0**	+1.6	**−7.3**	**+2.4**	**−5.7**	+1.5
wpq.doc	−33.3	—	−30.9	+1.8	−27.8	**+2.4**	−25.3	+2.0	−23.3	+1.6
wpq.path	−24.0	—	−21.6	**+2.0**	−20.5	+0.8	−19.5	+0.8	−16.7	**+2.4**
wpq.ost	−20.9	—	−20.0	+0.7	−19.2	+0.6	−17.4	+1.5	−13.9	+0.3
ran	−17.3	—	−19.1	−1.3	−18.8	+0.3	−18.3	+0.4	−17.6	+0.7

by the presence of nonrelevant information the Jeffrey's Conditioning and Binary Voting Models appeared most able to operate most effectively. The difference between all models was not significant with ANOVA across any iterations ($F(5, 54) = 1.844$, $p = .120$). Over time, all models increased precision slightly. With the exception of the *wpq.doc* model, all models took terms from relevance paths that extracted the most potentially useful parts of documents. While the documents were classified by the TREC assessors as nonrelevant, they had some features that made the SMART system rank them higher than other documents in the collection. They may contain additional words that could be of use in creating enhanced search queries.

Table X illustrates the marginal difference more clearly than Figure 5, showing the percentage change overall and the marginal percentage change at each iteration.

It should be noted that, using linear regression, there is no significant difference in the rate of learning in all models *after the first iteration* (*all* $r^2 \geq .8941$ and *all* $T(38) \geq 17.91$, $p \leq .05$). As was demonstrated in Scenarios 3a and 3b, the Jeffrey's Conditioning and Binary Voting Models performed better than the other models in the first iteration. When presented with paths from nonrelevant documents, these models seemed better able to extract useful terms. As shown in Table X, it was the first iteration that provided the overall increase in precision; after the first iteration the marginal changes were similar for all models.

5.4.2 *Relevance Learning.* We measured how well the implicit models trained themselves when given relevance information by the simulation. The relevance learning trend of the models was similar to Scenario 3, and was measured in the same way. Figure 6 shows the average Spearman and Kendall correlation coefficients across all 50 topics.

The results show that, in a similar way to Scenario 3, the models learned over time. However, since they were being shown information from nonrelevant documents they did not learn the relevant distribution (composed of relevant documents) at as fast a rate and did not finish with as high a correlation as in Scenarios 3a and 3b. The random model returned a coefficient value close to zero with both coefficients in 3a and 3b. However, in this scenario it was lower, suggesting it started at a low rate of learning and did not improve on this. The models based on *wpq* also performed poorly initially but improved gradually as the search proceeded.

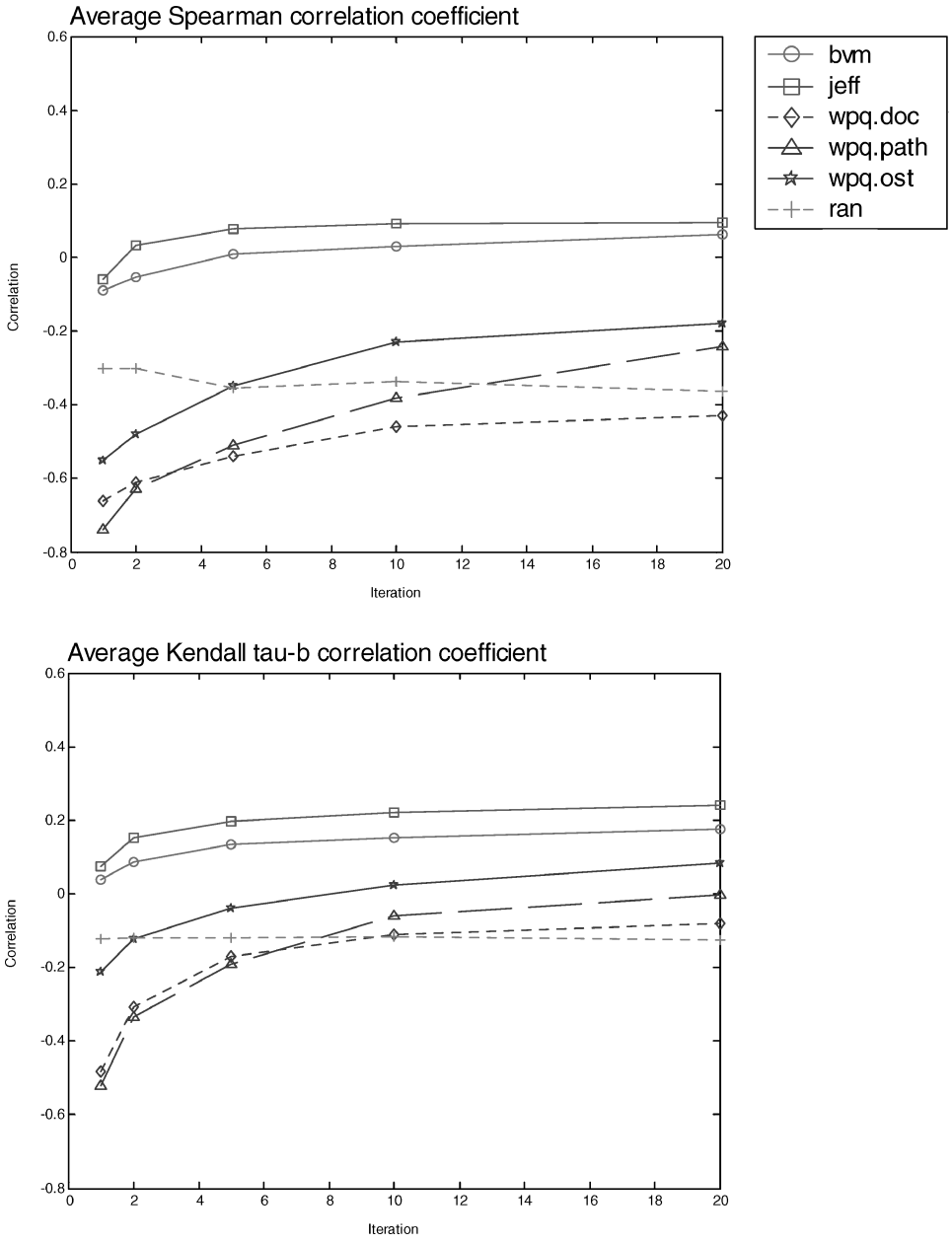


Fig. 6. Average correlation coefficient values across 10 experimental runs in Scenario 4a.

In a similar way to 4b, Scenario 3b revealed only a slight difference between the selection of paths randomly (as in 4a) and the use of the path length distributions. When paths were selected randomly, there was a restriction on their length, which could not exceed three steps. When the path length distributions were used, some paths were allowed to exceed this three-step boundary,

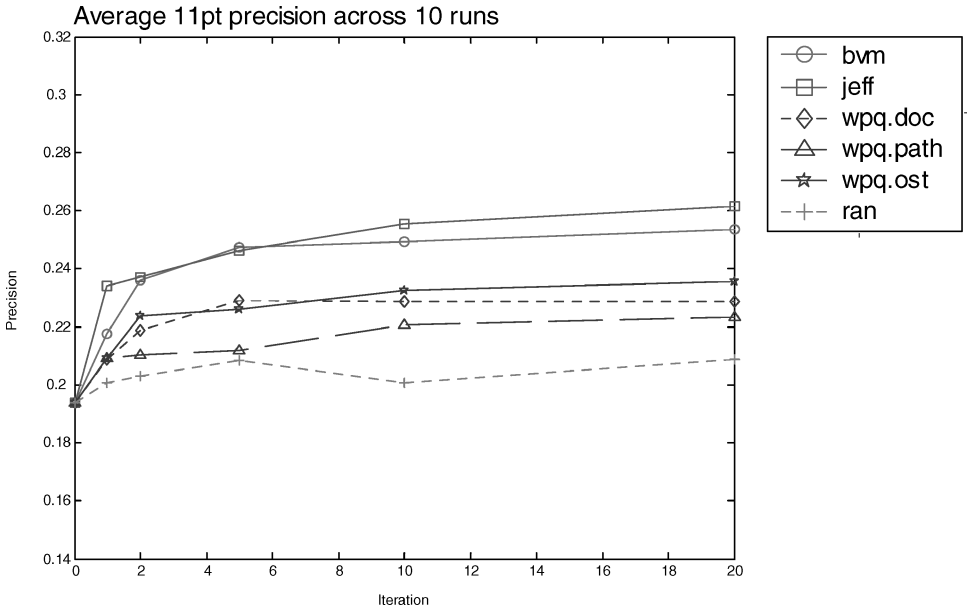


Fig. 7. Average 11-point precision across 10 experimental runs in Scenario 5a.

meaning the system was presented with more information. However, since this information was from irrelevant documents, it had a detrimental effect on the performance of all models and led to slightly larger reductions in search effectiveness.

5.5 Scenarios 5a and 5b: Related Paths

This scenario used the "Related Paths" approach described in Section 3.3.2 to select paths from relevant and nonrelevant documents. Search effectiveness (monitored through precision) and relevance learning (measured through correlation coefficients) were monitored for different levels of wandering. In this section we summarize the findings and present the average for all levels of wandering (i.e., the average for wandering levels at 10, 20, 30, 40, and 50%). This approach is potentially more realistic than the experimental scenarios presented so far in this article, as it is conceivable that searchers will view irrelevant information as they search.

5.5.1 *Search Effectiveness.* As in previous scenarios, the 11-point precision value was measured at iterations 1, 2, 5, 10, and 20. In Figure 7 we present the average precision value across all 10 runs and across all levels of wandering. The trend was the same as in earlier scenarios, with the Jeffrey's Conditioning and Binary Voting Models leading to overall increases in precision. However, because we introduced nonrelevant "noise" into the calculation, the overall increases in precision were not as large as in Scenarios 3a and 3b.

The percentage change in overall and marginal precision for each model is shown in Table XI.

Table XI. Percentage Change in Precision per Iteration in Scenario 5a (Overall change in first column, marginal change in second shaded column. Highest percentage in each column in bold.)

	Iterations									
Model	1		2		5		10		20	
bvm	10.9	—	17.9	**+7.8**	**21.7**	**+4.6**	22.3	+0.7	23.6	+1.7
jeff	**17.2**	—	**18.3**	+1.3	21.2	+3.6	**24.1**	+3.6	**25.9**	+2.3
wpq.doc	7.0	—	11.4	+4.7	15.3	+4.5	15.1	−0.2	15.3	+0.1
wpq.path	7.3	—	7.7	+0.5	8.5	+0.9	12.1	**+3.9**	13.1	+1.1
wpq.ost	7.3	—	13.3	+6.4	14.2	+1.0	16.6	+2.8	17.7	+1.4
ran	3.4	—	4.4	+1.0	7.0	+2.7	3.4	−3.9	7.1	**+3.9**

As the level of wandering increased, the increases in the level of precision decreased. Viewing information from nonrelevant documents (as Scenarios 4a and 4b demonstrate) was to reduce the overall effectiveness of all the term selection models. Nonetheless, the Jeffrey's Conditioning and Binary Voting Models still outperformed the others.

5.5.2 *Relevance Learning.* The models' ability to improve their understanding of what information is relevant was again measured using the Spearman and Kendall correlation coefficients. The values for both coefficients at iterations 1, 2, 5, 10, and 20 are presented in Appendices A.1 and A.2, respectively. Even though the models were shown potentially nonrelevant information, the results still demonstrate that the models were able to learn. However, their ability to do so was affected by the level of wandering. As wandering increased the rate at which the models learned relevance decreased.

In Scenario 5b, where path length distributions restricted the length of visited paths there were slight differences with this scenario. The restrictions imposed meant that the simulation had to choose paths that might not be as similar to the current path as other candidate paths, but had to be chosen to fill the percentage quota of the distribution. The overall effectiveness of the models was reduced by around 5% by imposing the path length restriction.

In the next section we discuss the results from this study, their implications for the design of search interfaces, and the development of techniques for the formative evaluation of such interfaces.

6. DISCUSSION AND IMPLICATIONS

In this article we have presented the evaluation of implicit feedback models using simulations that emulated the interaction of searchers. The implicit feedback models evaluated in this article *all* increased search effectiveness through query modification. However, two models performed particularly well: the Jeffrey's Conditioning Model and the Binary Voting Model. Both models improved precision and developed lists of terms that were closely correlated to those of the relevant distribution.

Initially, in most scenarios, the Jeffrey's Conditioning Model did not perform as well as the Binary Voting Model at the start of the search. However, after five paths it created more effective queries and from then on performed increasingly better than the latter. The Jeffrey's Conditioning Model used prior evidence

that was independent of the searcher's interaction. Initial decisions were made based on this prior evidence, and for the first few iterations it is reasonable to assume that this evidence still played a part in term selection. However, as more evidence was gathered from searcher interaction, the terms selected by the Jeffrey's Conditioning Model improved.

An advantage of the Binary Voting Model, and perhaps why it performed well in the initial stages, was that it did not rely on any prior evidence, selecting terms based only on the representations viewed by the searcher. However, the lists of potential terms offered stagnated after 10 paths; since in the Binary Voting Model the effect of the scoring was cumulative, the high-scoring, high-occurrence terms obtained a higher score after only a few initial paths, and could not be succeeded by lower-ranked terms in later paths. This often meant that the same query was presented in iterations 10 and 20.

The implicit feedback models learned relevance from the evidence provided to them by the simulation. This form of reinforcement learning [Mitchell 1997], where the model was repeatedly shown examples of relevant information, allowed us to test how well each model trained itself to recognise relevance. From the six models tested, our findings showed that the Jeffrey's Conditioning and Binary Voting Models learned at the fastest rate. In the first few iterations, those models based on *wpq* performed poorly in all retrieval scenarios, suggesting that these models need more training to reach an acceptable level of relevance recognition and that the Jeffrey's Conditioning and Binary Voting Models make a more efficient use of relevance information. Linear regression was used and compared the *rate of learning* against *precision* for each of the six implicit feedback models. The results showed that, for all models, the rate of learning (i.e., *Spearman's rho* and *Kendall's tau*) followed the same trend as precision (*all* $r^2 \geq .8154$ and *all* $t(38) \geq 5.34$, $p \leq .05$). The rate at which the models learned relevance appeared to match the rate at which they were able to improve search effectiveness.

The findings of the study show that the Jeffrey's Conditioning and Binary Voting Models were able to perform more effectively than the baselines when all the paths presented to them were from nonrelevant documents (Scenarios 4a and 4b) and when only a proportion of the paths were from nonrelevant documents (Scenarios 5a and 5b). While it is understandable that models can perform effectively when shown only relevant information, it is important for them to also perform well in situations where nonrelevant information is also shown. This is important in implicit, feedback models as they assume a degree of relevance in all the information searchers' views.

From the three models that implemented different versions of the *wpq* algorithm, the wpq.doc model performed best for all relevant documents (Scenarios 3a and 3b) and worst for all nonrelevant documents (Scenarios 4a and 4b). This model was more sensitive to the relevance of documents used than the path-based models. The document model must use all of the content of each document, whereas relevance paths comprise only the potentially useful parts of documents and hence reduce the likelihood that erroneous terms are selected. Since documents will typically be longer than relevance paths, the contribution

a single document makes to term scoring may typically exceed that of one relevance path.

In this study we have also shown that paths that lead to the largest marginal increases in relevance learning are those that are indicative of the term distribution they are trying to learn. That is, paths that are indicative of the terms that occur over all relevant documents are likely to be high-quality paths. There is no relationship between the number of steps in a path, the number of tokens in a path, or the percentage of stopwords in a path and the overall effectiveness of a path. Therefore, it is not how many words a path contains that determines the effectiveness of a relevance path, but what those words are, and how those words are distributed in the set of relevant documents.

For almost all the iterations on all the models, the marginal increases in precision and correlation reduced as more relevant information was presented. The models appeared to reach a point of saturation at around 10 paths, where the benefits of showing 10 more paths (i.e., going to iteration 20) were only very slight and perhaps outweighed by the costs of further interaction. It is perhaps at this point where searcher needs would be best served with a new injection of different information or explicit searcher involvement.

When employed in operational environments, the implicit feedback models should select good query modification terms regardless of the search topic. To test how the implicit feedback models performed for different topics, we conducted a topic-level analysis using each of the 50 TREC topics and examined how precision was affected by the topic used. In the analysis presented so far in this article, we have averaged our findings across all topics; now we present the results as an analysis aimed at identifying the extent to which topics influenced the performance of each implicit feedback model. To do this, we monitored precision values and computed the variability of search precision for all queries at each iteration. We do not present findings on a per query basis, but demonstrate how susceptible each model was to variations in the search topic. In each cell in Table XII, we show the variability of the precision (given by the average standard deviation as a percentage of the mean precision) for each query across all iterations.[15] We would expect model performance to be the same across all topics and therefore exhibit low variations in the precision values obtained. In situations where there was a high variance, there may well have been outlying queries for which there was very good or very bad performance.

A one-way independent measures ANOVA was used to test the significance of differences between queries. The results of this analysis suggest that for some scenarios there were some models with significant differences in the precision values (with $F(49,450)$ and $p < .05$). In situations where the ANOVA revealed significant differences, we applied Tukey's post hoc tests and found that in Scenarios 3 and 5—where relevant documents were used—certain TREC topics performed significantly better (e.g., topics 110, 125, 135, 150) or significantly worse (e.g., topics 109, 128, 148, 149) than most others. These topics shared

[15]No significant differences in variability *between* all five query iterations with a one-way repeated measures ANOVA, (all $F(4,245) \leq 1.85$, $p \leq .12$).

Table XII. Average Standard Deviation (as Percentage of the Mean) Across Five Iterations and 10 Experimental Runs for All Models and Scenarios (Cells with *no significant interquery differences in precision* are in bold.)

	Scenario					
Model	3a*	3b*	4a	4b	5a*	5b[a]
bvm	**30.30**	**34.20**	35.29	**33.91**	**32.08**	**31.03**
jeff	**32.15**	35.05	**30.12**	33.48	**33.89**	**33.38**
wpq.doc	41.54	44.42	43.62	44.07	42.97	42.16
wpq.path	43.94	43.18	40.46	41.93	44.84	42.92
wpq.ost	44.90	45.84	42.90	43.63	41.15	43.58
ran	74.82	71.63	72.63	70.13	73.18	71.06

[a]Scenarios that used only 43 of the 50 TREC topics.

no apparent attributes, and since this difference applied to all models and all scenarios it may be symptomatic of the document collection not supporting all search topics equally, either in the *volume* of information available or in the *quality* of information available. There was more interquery variation in the *random* model than in the other models since terms were not weighted sensibly and the performance of the query was dependent on quality of the terms selected.

The Jeffrey's Conditioning and Binary Voting Models were less dependent on the topic of the search query while leading to larger improvements in retrieval effectiveness over the other models. This suggests that these models are more robust, less dependent on the topic of the search query, and more useful for query modification. Query-level analyses of this nature can be used to test the robustness of RF algorithms. However, in this study only the topic of the query was varied and all queries were created in the same way (i.e., from the TREC topic title). It is conceivable that the models could be tested with specific or general queries, or searcher simulations used to mimic different query modification behaviors across a number of query iterations.

The same experiment was rerun using the *Wall Street Journal* 1990–1992 collection, accessible as part of the TREC initiative [Harman 1993]. This collection contains more documents than the SJMN collection and traditionally lends itself to smaller improvements in retrieval performance through query expansion. The same ranking of models was obtained with this collection as was obtained with SJMN. In future work, we will expand our simulations to use the TREC Web collections.

The Jeffrey's Conditioning Model performed best in all the scenarios in which the models were tested. This model is therefore a candidate for implementation in an experimental search interface, where its performance can be tested with human subjects and qualitative feedback on its performance obtained. The interface design evaluated in this study was developed separately and iteratively through user investigation. This article addresses what RF models are appropriate to support this user interface, not the other way round.

Simulation-based techniques of this nature can be useful for designers of search systems who can more fully test the suitability of implicit feedback models to the interface design and modify the models or interfaces where

appropriate. Through being able to test the interfaces without searchers, the costs of experimentation are reduced and the ability of the designer to develop more robust search interfaces is improved. Simulations of this nature can be used either after a prototype interface is built (as was the case in this study), or before the interface is built, to test its performance with every possible set of potential searcher interactions prior to development. This can assist system designers in identifying the strengths and weaknesses of the system (allowing them to eliminate interactions that could cause problems) and strengths and weaknesses of the RF algorithms (allowing them to choose a model that suits their needs).

The interaction modeled in this article assumed that all searchers in a scenario would interact in the same way. It is conceivable that a collection of simulated subjects could be assembled, each with a predetermined searching style. These searchers may have different ways of locating relevant information or different sets of relevance criteria when potentially relevant information is found. Determining what factors to vary, how to assemble and deploy the searchers, and running experiments form an intriguing challenge for IR researchers who use such simulations in the future.

In the next section, we present our conclusions.

7. CONCLUSIONS

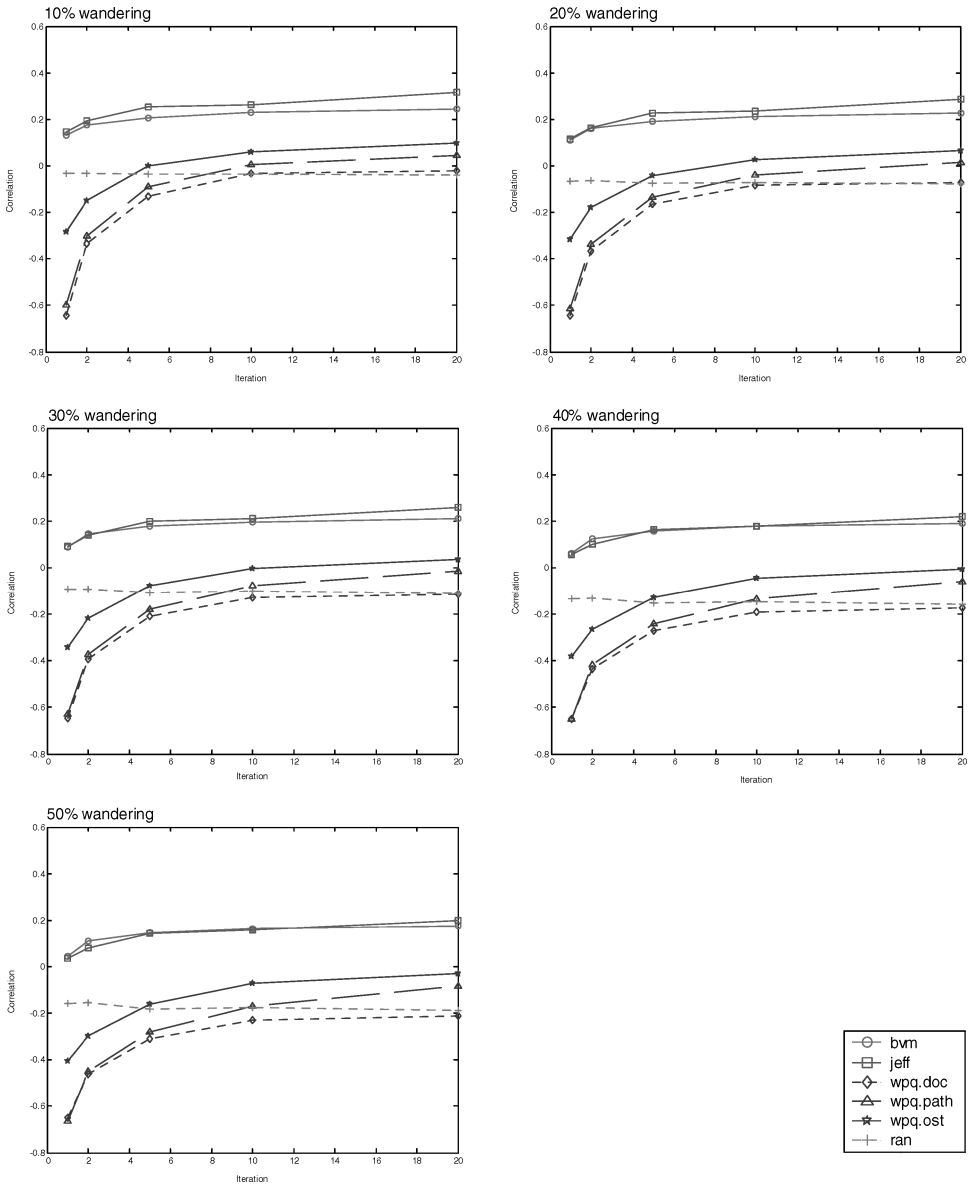
In this article we have presented a simulation-based evaluation to determine the effectiveness of a variety of implicit feedback models in predetermined retrieval scenarios independent of human subjects. These models depend on interaction with search interfaces as a source of evidence for the techniques they employ and use the exploration of the information space and the viewing of information at search interfaces as indications of relevance. Six implicit feedback models in total were tested, each employing a different term selection stratagem.

The simulated approach used to test the models assumed the role of a searcher "viewing" relevant documents and relevance paths between representations of documents. The simulation passed the information it viewed to the implicit feedback models, which used this evidence to select terms to best describe this information. We investigated the degree to which each of the models improved search effectiveness and learned relevance. From the six models tested, the Jeffrey's Conditioning Model provided the highest levels of precision, the highest rate of learning, and the highest levels of consistency across search topics. This model is therefore a candidate to be deployed in search interfaces and evaluated with human subjects.

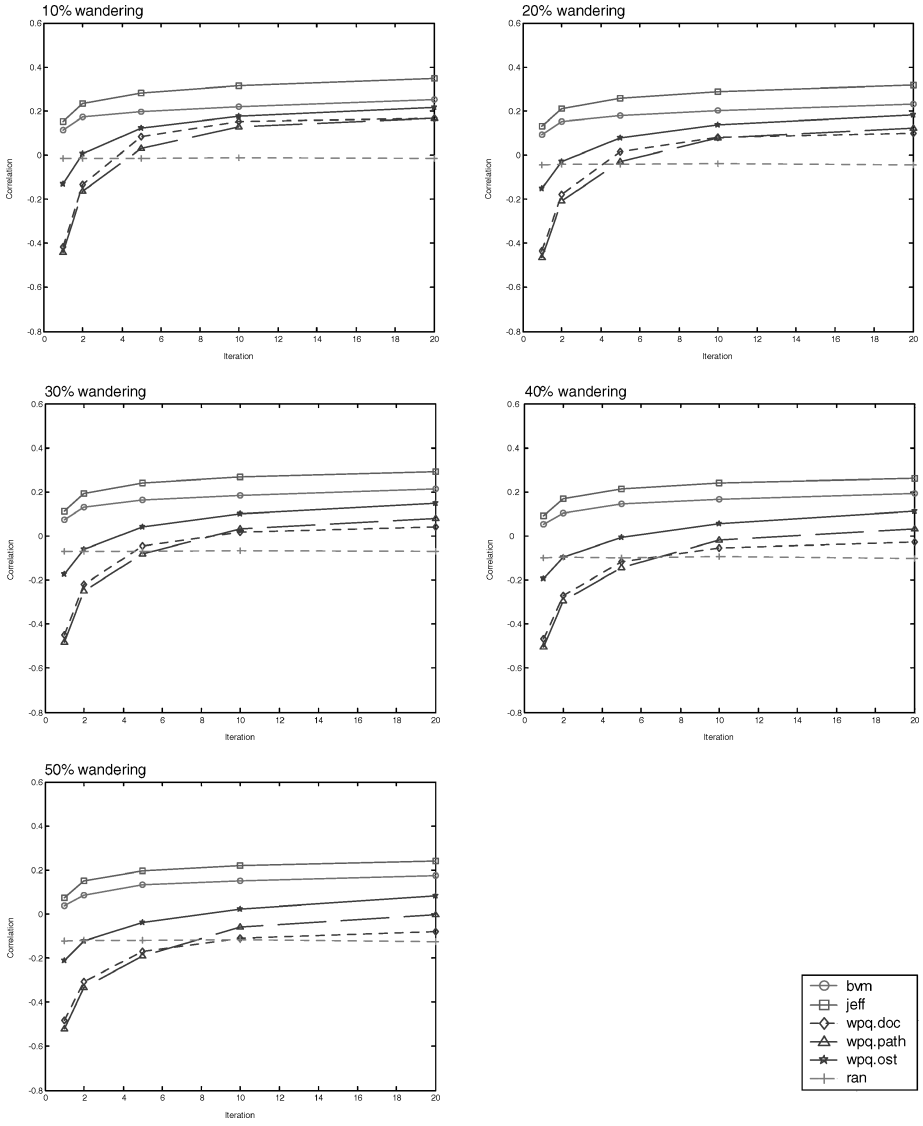
Since we used this methodology to evaluate implicit feedback models, not to evaluate the methodology itself, the conclusions we draw must be tentative for the moment. In future work we will evaluate the methodology through comparison with user-based evaluations,explore the development of more complete frameworks for IR evaluation based on searcher simulations, and explore the development of models of behavior to represent different situations, searchers, and searching styles.

APPENDIXES

A.1. Average Spearman Correlation Coefficient for Different Levels of Wandering



A.2. Average Kendall Correlation Coefficient for Different Levels of Wandering



REFERENCES

BARRY, C. L. 1998. Document representations and clues to document relevance. *J. Amer. Soc. Inform. Sci. 49*, 14, 1293–1303.

BORLUND, P. 2003. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Inform. Res. 8*, 3. Available online at http://informationr.net/ir/8-3/paper152.html R2.10.

BUCKLEY, C., SALTON, G., AND ALLAN, J. 1994. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. 292–300.

CAMPBELL, I. 2000. *The Ostensive Model of Developing Information Needs*. Unpublished doctoral dissertation. University of Glasgow, Glasgow, U.K.

CAMPBELL, I. AND 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*. 251–268.

CHALMERS, M., RODDEN, K., AND BRODBECK, D. 1998. The order of things: Activity-centered information access. *Comput. Netw. ISDN Syst. 30*, 1–7, 359–367.

CHI, E. H., PIROLLI, P., CHEN, K., AND PITKOW, J. 2001. Using information scent to model user information needs and actions on the Web. In *Proceedings of the Conference on Human Factors in Computer Systems*. 490–497.

CHI, E. H., ROSIEN, A., SUPATTANASIRI, G., WILLIAMS, A., ROYER, C., CHOW, C., ET AL. 2003. The Bloodhound Project: Automating discovery of Web usability issues using the infoscent simulator. In *Proceedings of the Conference on Human Factors in Computer Systems*. 505–512.

COSIJN, E. AND INGWERSEN, P. 2000. Dimensions of relevance. *Inform. Process. Manage. 36*, 4, 533–550.

FURNER, J. 2002. On recommending. *J. Amer. Soc. Inform. Sci. Tech. 53*, 9, 747–763.

HAMMING, R. W. 1950. Error-detecting and error-correcting codes. *Bell Syst. Tech. J. 29*, 147–160.

HARMAN, D. 1988. Towards interactive query expansion. In *Proceedings of the 11th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. 321–331.

HARMAN, D. 1993. Overview of the first TREC conference. In *Proceedings of the 16th Annual ACM SIGIR Conference of Research and Development in Information Retrieval*. 36–47.

HARPER, D. J. 1980. *Relevance Feedback in Document Retrieval Systems.* Unpublished doctoral dissertation. University of Cambridge, Cambridge, U.K.

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. *Inform. Process. Manage. 36*, 2, 207–227.

JEFFREY, R. C. 1983. *The Logic of Decision*. Chicago: University of Chicago Press, Chicago, IL.

KELLY, D. 2004. *Understanding Implicit Feedback and Document Preference: A Naturalistic User Study*. Unpublished doctoral dissertation. Rutgers University, New Brunswick, NJ.

KELLY, D. AND TEEVAN, J. 2003. Implicit feedback for inferring user preference. *SIGIR For. 37*, 2, 18–28.

LAM, W., MUKHOPADHYAY, S., MOSTAFA, J., AND PALAKAL, M. 1996. Detection of shifts in user interests for personalised information filtering. In *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. 317–325.

MAGENNIS, M. AND VAN RIJSBERGEN, C. J. 1998. The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*. 324–332.

MITCHELL, T. M. 1997. *Machine Learning*. McGraw-Hill, New York, NY.

MORITA, M. AND 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*. 272–281.

MOSTAFA, J., MUKHOPADHYAY, S., AND PALAKAL, M. 2003. Simulation studies of different dimensions of users' interests and their impact on user modeling and information filtering. *Inform. Retriev. 6*, 199–223.

PAEK, T., DUMAIS, S. T., AND LOGAN, R. 2004. WaveLens: A new view onto Internet search results. In *Proceedings on the ACM SIGCHI Conference on Human Factors in Computing Systems*. 727–734.

PIROLLI, P. AND CARD, S. 1995. Information foraging in information access environments. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 51–58.

ROBERTSON, S. E. 1986. On relevance weight estimation and query expansion. *J. Documentat. 42*, 182–188.

ROBERTSON, S. E. 1990. On term selection for query expansion. *J. Documentat. 46*, 4, 359–364.

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*. 213–220.

RUTHVEN, I., LALMAS, M., AND VAN RIJSBERGEN, C. J. 2003. Incorporating user search behavior into relevance feedback. *J. Amer. Soc. Inform. Sci. Tech. 54*, 6, 528–548.

SALTON, G. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, NJ.

SALTON, G. AND BUCKLEY, C. 1990. Improving retrieval performance by relevance feedback. *J. Amer. Soc. Inform. Sci. 41*, 4, 288–297.

SARACEVIC, T. 1975. Relevance: A review of and a framework for thinking on the notion of information science. *J. Amer. Soc. Inform. Sci. 26*, 6, 321–343.

SIEGEL, S. AND CASTELLAN, N. J. 1988. *Nonparametric Statistics for the Behavioural Sciences*. McGraw-Hill, New York, NY.

SPINK, A., GRIESDORF, H., AND BATEMAN, J. 1998. From highly relevant to not relevant: Examining different regions of relevance. *Inform. Process. Manage. 34*, 5, 599–621.

TOMBROS, A. AND SANDERSON, M. 1998. Advantages of query-biased summarisation in information retrieval. In *Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. 2–10.

VAN RIJSBERGEN, C. J. 1992. Probabilistic retrieval revisited. *Comput. J. 35*, 3, 291–298.

WHITE, R. W. 2004. *Implicit Feedback for Interactive Information Retrieval*. Unpublished doctoral dissertation. University of Glasgow, Glasgow, U.K.

WHITE, R. W. AND JOSE, J. M. 2004. A study of topic similarity measures. In *Proceedings of the 27th Annual ACM Conference on Research and Development in Information Retrieval*. 520–521.

WHITE, R. W., JOSE, J. M., AND RUTHVEN, I. 2003a. An approach for implicitly detecting information needs. In *Proceedings of the 12th Annual Conference on Information and Knowledge Management*. 504–507.

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. *Inform. Process. Manage. 39*, 5, 707–733.

WHITE, R. W., JOSE, J. M., AND RUTHVEN, I. 2004a. An implicit feedback approach for interactive information retrieval. *Inform. Process. Manage.* In press.

WHITE, R. W., JOSE, J. M., AND RUTHVEN, I. 2004b. Using top-ranking sentences to facilitate effective information access. *J. Amer. Soc. Inform. Sci. Tech.* In press.

WHITE, R. W., JOSE, J. M., VAN RIJSBERGEN, C. J., AND RUTHVEN, I. 2004c. A simulated study of implicit feedback models. In *Proceedings of the 26th Annual European Conference on Information Retrieval*. 311–326.

WHITE, R. W., RUTHVEN, I., AND 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*. 57–64.

WHITE, R. W., RUTHVEN, I., AND 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*. 93–109.

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 *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 249–256.

Received October 2004; revised March 2005; accepted April 2005