Using Searcher Simulations to Redesign a Polyrepresentative Implicit Feedback Interface

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

Abstract

Information seeking is traditionally conducted in environments where search results are represented at the user interface by a minimal amount of meta-information such as titles and query-based summaries. The goal of this form of presentation is to give searchers sufficient context to help them make informed interaction decisions without overloading them cognitively. The principle of *polyrepresentation* (Ingwersen, 1996) suggests that Information Retrieval (IR) systems should provide and use different cognitive structures during acts of communication to reduce the uncertainty associated with interactive IR. In previous work we have created *content-rich* search interfaces that implement an aspect of polyrepresentative theory, and are capable of displaying multiple representations of the retrieved documents simultaneously at the results interface. Searcher interaction with content-rich interfaces was used as Implicit Relevance Feedback (IRF) to construct modified queries. These interfaces have been shown to be successful in experimentation with human subjects but we do not know whether the information was presented in a way that makes good use of the display space, or positioned most useful components in easily accessible locations, *for use in IRF*. In this article we use simulations of searcher interaction behaviour as design tools to determine the most rational interface design for when IRF is employed. This research forms part of the iterative design of interfaces to proactively support searchers.

Keywords

Searcher simulations, interface design, implicit relevance feedback, polyrepresentation

1. Introduction

The principle of polyrepresentation (Ingwersen, 1996; Ingwersen & Järvelin, 2005) suggests that representations of different cognitive structures should be offered to searchers, and used by them during their interaction with an Information Retrieval (IR) system. The cognitive structures around which polyrepresentation is based are manifestations of human cognition, reflection or ideas. In IR they are typically transformations generated by a variety of human actors with a variety of *cognitive origins*. Author text, including document titles and their full-text are representations of cognitive structures intended to be communicated. These portions of text, have different *functional origins*; they have the same cognitive origin but were created in a different way or for a different purpose.

Polyrepresentative theory has generally been implemented through plausible inference techniques applied on networks of document representations (Turtle & Croft, 1990), or across networks of citations where those who cite documents have unique cognitive structures (Larsen & Ingwersen, 2002). While such research may benefit searchers, more work is needed to investigate the value of polyrepresentative principles in Interactive IR (IIR). Belkin *et al.* (1993) established that the polyrepresentative extraction of information needs is potentially more effective than eliciting the solitary, isolated query statements gathered by most IR systems. In a similar way, offering thesauri (Jones *et al.*, 1995) and clarification forms (Kelly *et al.*, 2005) during query formulation have been shown to lead to more effective query statements. Other techniques, such as Relevance Feedback (RF) (c.f. Salton & Buckley, 1990), where searchers indicate relevant documents to the system, are often under-utilised due to the additional cognitive burden they impose on searchers. Eliciting more expressive information need descriptions from searchers in these ways may improve retrieval but they depend on searchers' ability to explain their information problem, or impede on their search.

This article describes a stage in the development of a prototype search interface based on polyrepresentative principles. Interaction with this interface is used as Implicit Relevance Feedback (IRF) (Kelly & Teevan, 2003) to generate improved query statements that can be used to restructure or recreate search results. In Stage One we ran a comparative simulated study to choose the IRF model to underlie the interface (White *et al.*, 2005b). In Stage Two we ran a large study with human subjects to determine how interface support for the model we chose should be offered (White *et al.*, 2005a). In the study presented in this article (Stage Three) we use searcher simulations to help us make decisions about how the interface could be redesigned to make IRF

more effective. Interface (re)design is usually based on user involvement in the design process. User-centered design (Rubenstein & Hersh, 1984) is the most commonly adopted approach in IIR research; experimental participants are placed in a reacting role to experimental systems devised to address research questions (Muller *et al.*, 1993). However, user studies can be costly to arrange and run. In this article we demonstrate how simulations of some aspects of searcher behaviour can be used as tools to directly influence interface design decisions, potentially reducing the costs of IIR experimentation. Some of the simulations are tuned using interaction logs from the Stage Two study, and results are used to inform the construction of the most rational implicit feedback interface (i.e., the interface design deemed most effective when interaction with it is used for IRF).

The prototype interface used in Stage Two (shown in Figure 1) takes a searcher-provided query, submits it to a search system, retrieves the top-ranked documents, and extracts and dynamically creates sets of different document representations for presentation to the searcher. The interface implements a progressive revealment strategy where searchers can access an increasing amount of retrieved document content at the interface to help them decide whether they want to view its full-text. They do this by following interactive *relevance paths* between representations created from the same document; interacting with a representation guides the searcher to the next representation in the path or the full-text of the source document. Hovering over some representations with the mouse pointer, or clicking on icons next to others, prompts the interface to highlight the next step in the relevance path. Showing searchers progressively more information about a document to assist relevance assessment has already been used in related work (Zellweger *et al.*, 2000; Paek *et al.*, 2004). Clicking on the text of a representation takes searchers directly to the source document. The traversal of these paths is used by an IRF model to select terms for query expansion. The model used is based on Jeffrey's Rule of Conditioning (Jeffrey, 1983), and has been shown in the simulation-based study in Stage One to be more effective on our style of interface than other feedback algorithms (White *et al.*, 2004; 2005b).

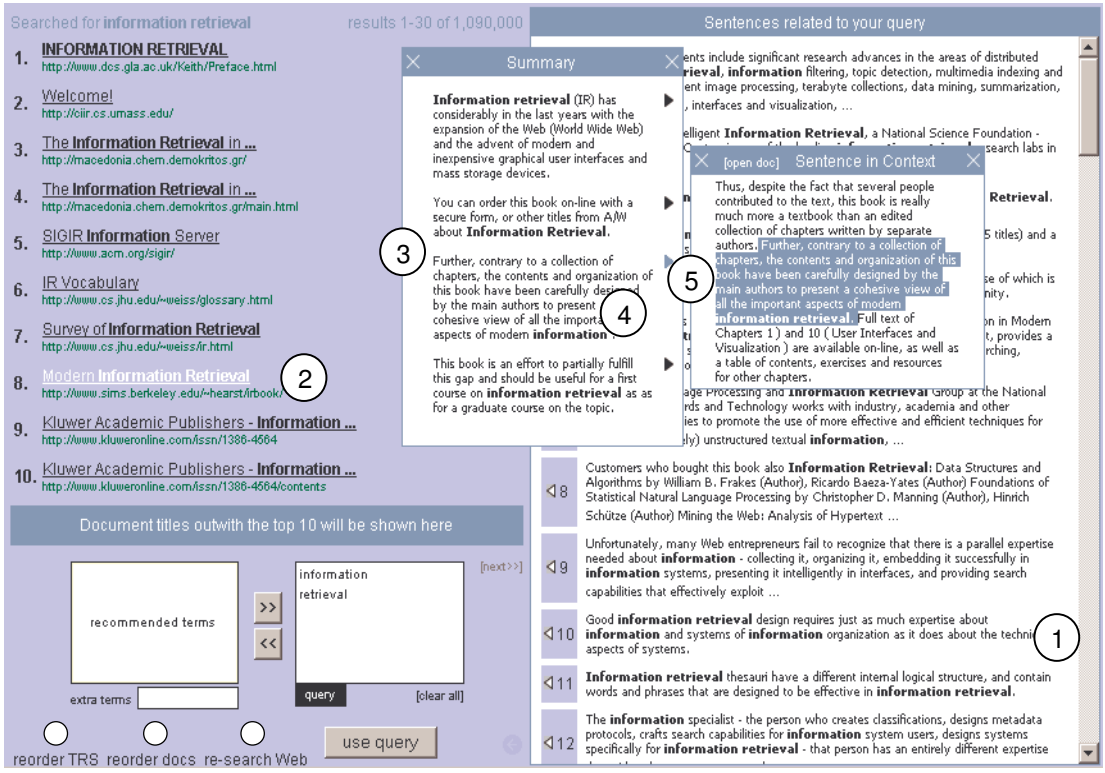

Figure 1. Prototype interface from Stage Two (White *et al.*, 2005a), which was deployed in a Web environment. The searcher has requested to see an extracted summary sentence in the context it appears in its source document.

To evaluate the components on this interface (i.e., representations and relevance paths) we need a method that is both: capable of providing objective evidence on their effectiveness, and allows us to vary and control how these components are used during the study to test them in a variety of ways. The richness of the potential set of interactions with this interface coupled with the fact that all searcher interaction is regarded as a source of RF, means that searcher simulations that extend the standard Cranfield model (Cleverdon *et al.*, 1966) to incorporate more interaction context may be a potentially robust way of determining component effectiveness (White *et al.*, 2005b). In the study we use a simulation-based approach to answer three questions with regard to the interface shown in Figure 1:

(i) What proportion of all information contained in top-ranked retrieved documents should be shown to searchers initially at the interface?

(ii) How many representations of each type should be shown at the interface?

(iii) What is the best arrangement of representations in relevance paths?

This list is by no means definitive, but represents the types of design-related questions that we were interested in answering.

The remainder of this article is structured as follows: Section 2 describes the simulations, including a description of the document representations used and the study itself. Section 3 describes and analyses study findings, and discusses how they affect the revised design of our interface. In Section 4 we discuss our methodology and findings, and conclude in Section 5.

2. Searcher Simulations

Typically, the only searcher interaction simulated in standard IR experimentation – such as the *ad hoc* task at TREC (Voorhees & Harman, 2005) – is the provision of queries. In RF experiments the interaction simulated is generally both the submission of queries and the provision of RF through marking relevant documents over a series of feedback iterations (Buckley *et al.*, 1994). However, our interface, shown in Figure 1, facilitates more interaction than these forms of result presentation, and assumes that every interaction is an indication of relevance; a different evaluation strategy is therefore required. Simulation-based methods have been used in previous studies to test query modification techniques (Harman, 1988; Magennis & Van Rijsbergen, 1998; Ruthven, 2003; White *et al.*, 2005b), or to detect shifts in the interests of computer users (Lam *et al.*, 1996; Mostafa *et al.*, 2003). These methods are worthwhile since: (i) they are less time consuming and costly than experiments with human subjects, (ii) they allow the comparison of IR techniques in different retrieval scenarios, and (iii) they 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 simulations in a different way to previous studies: to simulate searcher interaction with a variety of document representations (or relevance paths) at the results interface, use this interaction as IRF, and monitor the effectiveness of the resultant expanded queries to influence design decisions.

The simulation assumes the role of a searcher, browsing the results of an initial retrieval. The information content of the relevant documents in the top-ranked documents in the first retrieved document set constitutes the information space that the searcher must explore. We use relevant documents as the source for the representations simulated searchers interact with since they are more likely to be topically coherent than the non-relevant documents. All interaction simulated was with these documents and a new information space is never generated. This allowed us to

evaluate the contribution of each of the representations as sources of IRF between searcher-defined query iterations. In the simulation searchers were modelled using different strategies: (i) vary the proportion of all representations interacted with, (ii) vary the amount of interaction with each type of representation, and; (iii) vary the arrangement of representations in relevance paths. Based on simulated interaction the IRF model generates a new query composed of the original query and the top six query expansion terms.[1] This revised query is used to retrieve a new document set that is then scored using mean average precision, a standard IR evaluation metric.

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 & Ingwersen, 2000; Borlund, 2003). However, they can allow for a more complete analysis of the retrieval effectiveness of components and help decide which should be given prominence in our revised interface. In the remainder of this section we provide more information about the document representations used and describe the experimental methodology.

2.1 Document Representations

Many of the representations used in this simulated study are sentence-based, including two that comprise only individual sentences. Ingwersen (1996) suggests that paragraphs are the smallest semantically confined unit of a document that can effectively be used in any application of polyrepresentative principles. Paragraphs have been used as passage-level evidence for the indexing and subsequent retrieval of documents (Callan, 1994). However our previous experience with searchers has shown that sentences may contain the information necessary to satisfy their information needs, or may provide a means through which they can access relevant documents (White *et al.*, 2003b). Allowing relevance assessments at the sentence level also allows for more precise feedback on what information meets searcher needs. The following are considered to be representations of a document:[2]

1. **Top-ranking sentences (TRS)** are the highest scoring query-relevant sentences extracted from top documents at retrieval time. A maximum of four sentences per document are extracted and these are generally presented to searchers as a list, ranked based on their sentence score and independent of the ranking of their source documents. TRS are selected based on the summarisation approach described in White *et al.* (2003a).

[1] This number of query terms has been shown to be effective in previous query expansion studies (Harman, 1988).
[2] The numbers correspond to those used in Figure 1.

2. **Document titles** are typically short and express the main themes of a document.
3. **Document summaries** are based on the query devised by the searcher and comprise the best four TRS from a single document ranked in sentence score order.
4. **Individual sentences in document summaries**.
5. **Summary sentences in document context** (i.e., sentence from document summary plus preceding and following sentence in the source document) are also available for searchers to view. Sentences in context are created immediately after the retrieved documents have been summarised (i.e., after query submission and before result presentation). These can be of particular use when a sentence is *anaphoric* i.e., refers back to a previous sentence in the document, or *cataphoric* i.e., refers forward to a forthcoming sentence in the document. For example, given the two sentences: "Alexander Graham Bell invented the telephone. He was born in Scotland and emigrated to Canada when he was just 23 years old." The pronoun "he" in the latter sentence is referent to the "Alexander Graham Bell" in the former sentence. This is an anaphoric reference and can cause ambiguity if the latter sentence is shown without the former. Problems with anaphoric and cataphoric references are symptomatic of the extractive approach we use to select sentences. Although the resolution of such problems has been addressed previously in IR research (Liddy, 1990; Paice & Jones, 1993), these approaches have not always been successful. Our approach employs no algorithmic support, but showing sentences in the original document context may contribute somewhat to the resolution of these issues by searchers.

There is intentional redundancy in the representations that searchers interact with at our interface. A single sentence may appear in four of the five document representations: in the list of TRS, in the document summary, as a summary sentence, and as a sentence in context. Searcher interaction with the same sentence in a number of representations reinforces the relevance of the terms contained in that sentence.

The sentence-based representations (i.e., TRS, document summaries, summary sentences, and sentences in context) have different functional origins and the same cognitive origins (different from the author of the source document). The representations are created using algorithms devised by system designers and are selected based on queries submitted by a searcher, both cognitive agents.

2.2 Relevance Paths

Relevance paths connect document representations at the interface. In the prototype interface these representations were arranged in a particular order based on the intuition of the system designer. This order is shown in Figure 2. Searchers can begin relevance paths at either TRS or titles, and path length can range from one to five representations.

TRS Title Summary Summary Sentence Summary Sentence in Context.

Figure 2. Relevance path ordering in prototype interface (numbers correspond to Figure 1).

A simulation-based approach allows us to test many possible path orderings and determine the best arrangement for use in a revised interface design. There are 54 possible relevance path routes between representations in the interface used in Stage Two. That is, there are 54 different combinations of five and fewer representations in the order they appear in Figure 2.

2.3 System, Corpus and Topics

The flexible Terrier IR platform[3] (Ounis *et al.*, 2005) was used in the experiment to index and search the corpus using a best-match retrieval algorithm. Terrier was used due to the ease with which document collections could be indexed and searched. Two test collections were used: the San Jose Mercury News (**SJMN** 1991) and the Wall Street Journal (**WSJ** 1990-1992) document collections taken from the TREC initiative (Voorhees & Harman, 2005). These collections have been used successfully in previous experiments of this nature (Ruthven, 2003; White *et al.*, 2005b). Relevant summary statistics of these collections are shown in Table 1.

Table 1. Test collection statistics.

Collection	SJMN	WSJ
Total number of documents	90257	74520
Mean average document length (including title)†	410.7	388.2
Relevant documents per topic	55.6	30.3

† Document length is measured in words, including stopwords

The WSJ collection contained on average fewer relevant documents per topic than the SJMN collection. This contributed to making improvements in retrieval effectiveness through query expansion potentially more difficult on the WSJ collection.

[3] http://ir.dcs.gla.ac.uk/terrier/

The topics are assumed to represent the information goal of the simulated searcher that is refined during the search. 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 is appropriate because it is similar in length and content to real searcher queries. The simulation retrieves the top 30 documents[4] for each of the 50 TREC topics used as queries in this study; these results can contain both relevant and non-relevant documents. In our study runs require relevant documents. However, for some topics, there are no relevant documents in the top 30 results, making the execution of some scenarios problematic. We exclude topics with no relevant documents in the initial top 30 results from the experimental runs. Table 4, later in this article, shows the number of topics that were used in each experimental run.

2.4 Measures

The primary measure used in this study is *mean uninterpolated average precision* (MAP), the mean average of the average precision values obtained across all topics. Precision measures the proportion of relevant documents in all documents retrieved. We assume that searchers are interested in maximising this value and that it gives a good estimation of the utility of an expanded query.

2.5 Methodology

The simulation is fully automated and after initiation requires no input from the experimenter until the experimental run (or set of runs) is complete. The following methodology is used by our simulation:

For each TREC topic use title field to form query

a. Submit query to the Terrier IR system.
b. For each relevant document in the top 30 documents retrieved:
    i. Get full-text of document;
    ii. Extract and score all sentences, and;
    iii. Create all document representations.
c. Form pools of all representation types (e.g., all TRS, all titles), and all relevance path arrangements.

[4] This is the number of documents that are retrieved and summarised in response to searcher queries in the prototype interface. Limiting this number to 30 allows the interface to respond to searchers in a timely manner.

d. Select and use representations from these pools based on the scenario described in the next section.

2.6 Scenarios

In this section we describe the scenarios that approximate searcher behaviour in a variety of ways. This scenario set is not exhaustive, but does allow us to address our three research questions: what proportion of all information to show? How many representations of each type to show? What is the best arrangement of representations in relevance paths? We now describe the three scenarios in more detail.

2.6.1 Scenario 1: Proportion of all representations

There is a trade-off in interface design between showing searchers sufficient information to make interaction decisions and solve problems, and overloading them cognitively, potentially hindering their interaction. This scenario varies the proportion of all representations that simulated searchers interacts with and tries to answer the question: *how much information should we show initially at the results interface?* An initial retrieval yields 30 documents that are used to create a maximum of 420 document representations comprising 120 TRS, 30 titles, 30 summaries, 120 summary sentences and 120 sentences in context. Showing all of these representations to searchers at the interface would not be practical. In this scenario we vary the number of representations that the simulated searcher interacts with (i.e., 1, 5, 10, 20, 50, and 100 representations), use interaction with these representations as a source of IRF, and investigate the impact of different amounts on retrieval effectiveness. In earlier research, we suggested that there was a "saturation point" in IRF beyond which further interaction would not substantially improve the expanded queries generated (White *et al.*, 2004). Scenario 1 is aimed at determining this point for use in interface redesign. This scenario differs from the other scenarios described in this section since it takes two forms: (A) use some proportion of all information, randomly selected, (B) use some proportion of all information selected based on relatedness to previously viewed information (determined by a Cosine similarity measure). The representation from which to start the simulated interaction is randomly selected for each topic from the pool of relevant representations.

2.6.2 Scenario 2: Amount of interaction with each type of representation

This scenario varies the amount of interaction with each of the five types of document representations to answer the question: *how much of each representation type should we show at*

the interface? We compute the average marginal effect on MAP of each representation type. This helps us establish which of the representation types are most effective for IRF on average, and hence which should be given the most space on the new version of the interface and made most easily accessible for the searcher. This scenario uses all representations of each type.

2.6.3 Scenario 3: Arrangement of representations in relevance paths

The order in which feedback is provided can affect both the performance of RF algorithms (Salton & Buckley, 1990) and searcher feelings of satisfaction with the RF system (Tianmiyu & Ajiferuke, 1988). The effectiveness of RF is dependent on the information provided as feedback during the current feedback iteration and the feedback that has come before. The arrangement of representations in relevance paths is important and can determine the retrieval effectiveness of the expanded query created by the IRF models. Representation arrangement in the relevance paths in our prototype interface was derived from our informed intuition about what arrangement was appropriate. Simulations provide us with a mechanism to experiment with a variety of arrangements at relatively low cost. This scenario tries to answer the question: *what is the best arrangement of representations in relevance paths?* It tests different path arrangements to see whether the path order we selected in our interface can be improved. The path arrangements of interest in this study are shown in Table 2. Each row in the table contains the representations in the order they will occur in the path (from left to right). The maximum number of representations of each type is shown below the symbol for each representation. This scenario uses all generated paths of each arrangement.

Table 2. Relevance path arrangements.

Arrangement	Number of paths
T (4) → I (1) → U (1) → S (4) → C (1)	16
T (4) → I (1) → U (1) → S (4)	16
T (4) → U (1) → S (1) → C (4)	16 †
T (4) → U (1) → S (4)	16 †
T (4) → I (1) → U (1)	4
T (4) → I (1)	4
T (4) → C (1)	4 †
T (4)	4
T (4) → U (1)	4 †
I (1) → U (1) → S (4) → C (1)	4
I (1) → U (1) → S (4)	4
U (1) → S (4)	4
U (1) → S (4) → C (1)	4
U (1) → T (1) → C (4)	4 †
I (1) → U (1)	1
I (1)	1
U (1) → T (1)	1 †
U (1)	1 †
Total	108

T = Top-Ranking Sentence, I = Title, U = Summary, S = Summary Sentence, C = Sentence in Context
† This relevance path arrangement does not exist on the prototype interface from Stage Two.

Not all possible path arrangements are listed in Table 2. Those that are listed represent the union of the set of orderings that our experience with the interface and IRF model suggests might be worth using, and the set that subjects in previous studies (White *et al.*, 2005b; White *et al.*, 2006) had suggested might be useful. This demonstrates how system designers can be selective in how they use simulations to address issues that interest them; they do not have to simulate all possible combinations of a particular interaction behaviour unless the situation demands it. Comparing the results of this scenario with the results of Scenario 2 allow us to determine the value of the relevance path as an interaction technique to support information seeking.

2.7 Experimental Runs

The experiment contains four experimental runs (shown in Table 3). These runs vary two factors: how relevance paths / representations are selected and the number of relevance paths / representations presented as RF.

Table 3. Experimental runs.

Scenario	Test	Experimental variables		
		Selection	**Number of paths or representations**	**Average execution time (mins)**
1	A	Random	1,5,10,20,50,100	31
	B	Similarity	1,5,10,20,50,100	43
2	–	–	All of each representation type	85
3	–	–	All of each path arrangement	162

In Scenario 1 the effect of IRF is cumulative, there are 10 runs for each of the two tests (A and B), and we report on the MAP at predetermined step values, averaged across all runs. There is no need to do this for Scenarios 2 and 3 as no selection of a subset of available representations or relevance paths takes place. No special system requirements were needed to perform the runs,[5] and the average execution time for each scenario is shown in the last column of Table 3.

2.8 Incorporating Interaction Logs

The interaction simulated in Scenario 1 represents our intuition about how searchers may interact with the interfaces we developed based on our experience in running prior user studies with them. To improve the realism of the simulations we used the interaction logs generated by 48 subjects using the interface for one hour of searching in the Stage Two user study (White *et al.*, 2005a). The interaction logs are used to determine the number of representations of each type selected in Scenario 1 based on the proportion of interaction with each type during the user study. Table 4 shows the representation types and the proportion of all interaction that was recorded with each of them.

[5] A desktop PC with a reasonable specification – Pentium 4 (2.8GHz), 1GB RAM, 80 GB hard disk – was used.

Table 4. Proportion of all interaction with each representation type in Stage Two user study.

Representation type	Proportion (%)
TRS	21.32
Title	32.41
Summary	19.32
Summary Sentence	15.38
Sentence in Context	9.77

The values in Table 4 determined the number of representations of each type used in Scenario 1. For example, if 10 representations were to be selected in accordance with the scenario then two would be top-ranking sentences, three would be titles, two would be document summaries, two would be summary sentences, and one would be a sentence in its original document context. Using interaction logs in this way allows us to make better choices about setting parameters in the simulation.

In the next section we present the findings of our study and relate each of the findings directly to interface design decisions.

3. Findings and Design Implications

In this section we describe the findings of the study, grouped by research question. Nonparametric statistical testing (Siegel & Castellan, 1988) is used where appropriate and $p < .05$ unless otherwise stated. To test the effectiveness of our approach we apply our findings to justify changes to the interface described in the introduction to this article. In Figure 3 we present a schematic of this interface prior to running this simulated study.

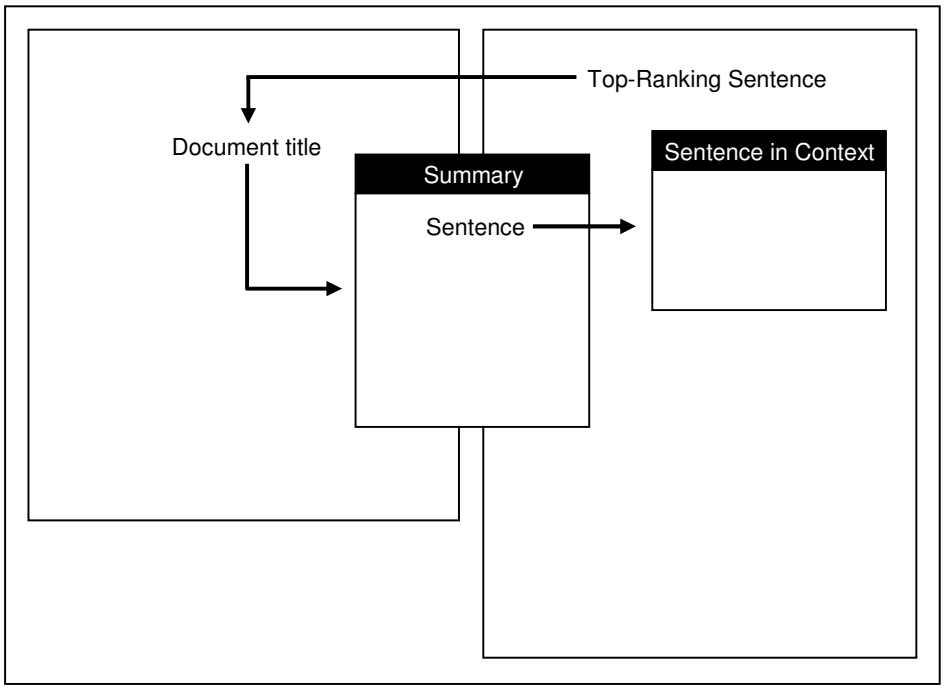

Figure 3. Interface schematic prior to this study (with one full relevance path marked).

In this section we relate findings from the study directly to this interface design, and generate a revised schematic based on experimental outcomes. The three aspects we focus on are: amount of information displayed, accessibility, and arrangement of representations in relevance paths.

3.1 Amount of information displayed (Scenario 1)

The first issue we address is how much information should be shown to searchers initially at the results interface. To determine how much information to show, we devise simulations that select document representations randomly, treating each one as relevant (Test A), and others that select representations based on their relatedness to previously viewed information (Test B). Using the simulations to approximate searcher interaction with up to 100 representations in each case we seek to determine whether there is a point at which making any more information available for display does not lead to significant improvements in IRF performance. The two tests were each run 10 times for each collection and the average MAP after a predetermined number of steps are shown in Table 5. As can be seen, the number of topics (shown in column 3) shrinks and the baseline precision value changes as the number of representations grows. This is since not all topics have sufficient relevant documents in the top-ranked results to construct large numbers of representations. The baseline MAP is computed as the mean average precision value obtained

from initial best-match retrieval for the topics that yielded any relevant documents in this retrieval (42 topics for the SJMN collection and 46 topics for the WSJ collection).

Table 5. Average MAP across different numbers of document representations.

SJMN						
Test	**Step**	**Topics**	**Baseline MAP**	**Avg. MAP**	**Change (%)**	**Marginal change (%)**
A	1	42	.2223	.2490	+ 12.01	n/a
A	5	42	.2223	.2850	+ 28.21*	+ 16.20
A	10	42	.2223	.3231	+ 45.34**	+ 17.13
A	20	38	.2349	.3677	+ 56.56**	+ 11.22
A	50	30	.2644	.3864	+ 46.15***	– 10.41
A	100	25	.2755	.4098	+ 48.75***	+ 2.60
B	1	42	.2223	.2447	+ 10.08	n/a
B	5	42	.2223	.3040	+ 36.75**	+ 26.67
B	10	42	.2223	.3690	+ 65.99**	+ 29.24
B	20	38	.2349	.3944	+ 67.93***	+ 1.94
B	50	30	.2644	.4171	+ 57.76***	– 10.17
B	100	25	.2755	.4212	+ 52.89***	– 4.87
WSJ						
Test	**Step**	**Topics**	**Baseline MAP**	**Avg. MAP**	**Change (%)**	**Marginal change (%)**
A	1	46	.2264	.2312	+ 2.12	n/a
A	5	46	.2264	.2434	+ 7.51	+ 5.39
A	10	46	.2264	.2533	+ 11.88*	+ 4.37
A	20	44	.2238	.2643	+ 18.10**	+ 6.22
A	50	33	.1932	.2219	+ 14.86**	– 3.24
A	100	20	.1643	.2028	+ 23.43***	– 8.57
B	1	46	.2264	.2355	+ 4.02	n/a
B	5	46	.2264	.2554	+ 12.81*	+ 8.79
B	10	46	.2264	.2743	+ 21.16**	+ 8.35
B	20	44	.2238	.2844	+ 27.08***	+ 5.92
B	50	33	.1932	.2329	+ 20.55***	– 6.53
B	100	20	.1643	.2241	+ 36.40***	+ 15.85

Wilcoxon Signed-Rank Test: * $p < .05$, ** $p < .01$, *** $p < .001$

Wilcoxon Signed-Rank Tests were performed to determine the significance of the differences between the average MAP values after varying degrees of interaction and the baseline MAP value normalised for the number of topics. The significant differences are indicated in Table 5. The MAP values are slightly higher in Test B since representations were selected based on their relatedness to previously "viewed" information, whereas in Test A representations were selected at random. Test B interaction was therefore potentially more coherent than Test A interaction.

Simulations generally performed worse on the WSJ collection than the SJMN collection.[6] Given the fewer relevant documents in this collection this was to be expected. The findings of this analysis reveal that there may be a point between 20 and 50 representations where more interaction does not necessarily yield better IRF performance. A conservative decision based on this finding is that the IRF model does not benefit from the searcher interacting with more than 50 document representations. Based on this we opt to show no more than 50 document representations at the revised interface at any given time. However, since there are 5 types of representation, we must decide what proportion of each representation type to display, and how accessible to make the representations given limited display space. In the next section we address this issue.

3.2 Accessibility (Scenario 2)

It is prudent if all interaction is to be used as some form of IRF to make those representations that are going to yield the most effective queries when taken as feedback most accessible to searchers. To do this we evaluate the marginal MAP of the 5 types of representation independently, then devote the most display space and/or make most accessible, those representations that score highest. Table 6 shows the mean average precision across all useable topics (42: SJMN, 46: WSJ). The average percentage change is shown from an initial value of .2223 for the SJMN collection and .2264 for the WSJ collection. The total number of representations of each type that were used to compute the average MAP are shown in the columns marked "Num."

Table 6. Mean average precision by representation type.

Representation	SJMN			WSJ		
	MAP	Num.	Change (%)	MAP	Num.	Change (%)
Title	.2898	1532	+ 30.36***	.2548	1294	+ 12.54*
Summary	.2705	1532	+ 21.68***	.2443	1294	+ 7.91
Sentence in Context	.2578	6128	+ 15.97**	.2344	5176	+ 3.53
Summary Sentence	.2407	6128	+ 8.28*	.2314	5176	+ 2.21
Top-Ranking Sentence	.2396	6128	+ 7.78*	.2291	5176	+ 1.19

Wilcoxon Signed-Rank Test: * $p < .05$, ** $p < .01$, *** $p < .001$

Titles, summaries and sentences in context yielded the largest increase in retrieval effectiveness when used for IRF. In the current interface design the summaries and summary sentences are less

[6] Although there were no significant differences between tests A and B for either collection and for all steps (Mann-Whitney Tests, $ps \geq .132$, $\alpha = .01$). A Bonferroni correction was used to control the experimental wise error rate, i.e., .05 divided by the number of pair-wise comparisons.

accessible than the TRS, which occupy half of the available display space (see Figure 1), but do not lead to large improvements in retrieval effectiveness when used for IRF. We applied a Kruskal-Wallis test to the data and the findings indicated significant differences between the five representation types ($\chi^2(4) \geq 12.86$, $\underline{p}s \leq .012$). Further analysis using Dunn's *post hoc* tests revealed statistically significant differences between some of the representation types. Table 7 shows the outcome of this post-hoc testing for every representation type pair for each collection, significant differences are shown in bold.

Table 7. Dunn's post-hoc tests among representation types in SJMN and WSJ collections.

Representation	TRS	Title	Summary	Summary Sentence	Sentence in Context
TRS	–	**.0165**	**.0120**	.3323	.1264
Title	**.0071**	–	.1430	**.0211**	**.0154**
Summary	**.0094**	.0894	–	.1481	.2003
Summary Sentence	.2600	**.0160**	**.0349**	–	.4619
Sentence in Context	**.0476**	**.0233**	**.0479**	.1331	–

SJMN collection, WSJ collection (italicised)

The results of this analysis suggest that document titles and summaries lead to significantly better queries than the other representation types when used for IRF. This suggests that they should be given more display space and made more accessible at the interface. TRS and sentences in context do not appear to be as useful for IRF. This is perhaps because they are shorter than a four-sentence document summary and potentially less informative than a document title. Participants in many studies (White *et al.*, 2004; 2005a,c) have remarked how useful the sentences can be to facilitate exploration. However, in this article we are measuring their usefulness *from the system's perspective* when used for IRF. Therefore, despite previous study participants' opinions on their utility we will give the TRS a less predominant role in the revised interface that emerges from this study.

3.3 Relevance Path Arrangements (Scenario 3)

There were 18 possible relevance path arrangements that we were interested in studying. From these arrangements we were searching for the most effective arrangement of representations in terms of the quality of the expanded query produced based on interaction with it. One of the powerful features of using simulations of searcher behaviour is that many alternatives can be explored in a short time with limited expense for the experimenter. Table 8 shows the ordering of the relevance paths ranked based on the average MAP across 10 experimental runs. The

"Arrangement" column displays the order in which representation types occur in the relevance path (e.g., "UIC" represents: "Summary *then* Title *then* Sentence in Context"). The percentage changes shown in Table 8 are from the baseline MAP for each of the collections (i.e., SJMN: .2223, WSJ: .2264).

Table 8. Mean average precision by path arrangement (sorted by SJMN MAP).

Arrangement	SJMN				WSJ			
	MAP	Num.	Change (%)	*R*	MAP	Num.	Change (%)	*R*
UIC	.3445	6128	+ 54.97***	1	.2945	5176	+ 30.08***	1
UI	.3421	1532	+ 53.89***	2	.2921	1294	+ 29.02***	2
U	.3236	1532	+ 45.57***	3	.2736	1294	+ 20.85**	3
TU	.3115	1532	+ 40.13***	4	.2715	1294	+ 19.92**	4
USC	.2967	6128	+ 33.47***	5	.2671	5176	+ 17.98**	5
US	.2859	6128	+ 28.61***	6	.2654	5176	+ 17.23*	6
I	.2818	1532	+ 26.78**	7	.2573	1294	+ 13.64*	7
IU	.2673	1532	+ 20.25**	8	.2548	1294	+ 12.54	8
TC	.2586	6128	+ 16.35**	9	.2450	5176	+ 8.22	9
TUSC	.2454	6128	+ 10.39*	10	.2435	5176	+ 7.55	10
TUS	.2451	24512	+ 10.27*	11	.2384	20704	+ 5.30	13
TIUSC	.2438	24512	+ 9.68	12	.2424	20704	+ 7.07	12
TIUS	.2394	24512	+ 7.69	13	.2313	20704	+ 2.16	14
IUSC	.2387	6128	+ 7.38	14	.2431	5176	+ 7.38	11
IUS	.2353	6128	+ 5.86	15	.2148	5176	– 5.12	15
TIU	.2332	6128	+ 4.92	16	.2142	5176	– 5.39	16
TI	.2284	6128	+ 2.76	17	.2080	5176	– 8.13	17
T	.2284	1532	+ 2.75	18	.2075	1294	– 8.35	18

T = Top-Ranking Sentence, I = Title, U = Summary, S = Summary Sentence, C = Sentence in Context
Wilcoxon Signed-Rank Test: * $p < .05$, ** $p < .01$, *** $p < .001$
R = rank

We are interested in the relative differences rather than actual differences between the MAP values that the relevance path arrangements obtained on the SJMN and WSJ collections. For this reason we calculated the non-parametric Kendall's *tau-b* correlation coefficient. The resultant value was .92, indicating a strong correlation between the two rankings. Most importantly for our interface design, the same path arrangements occupied the top ten positions on each collection. Document summaries appear to play a part in the most successful relevance path routes. This finding, coupled with those from earlier scenarios suggest that titles and summaries should play a prominent role in the revised interface, that relevance paths should generally be short (i.e., contain up to 3 representations), and that relevance paths should include at least a document summary.

The average MAP values in highly-ranked relevance paths (Table 8) are generally higher than those for a similar number of representations not linked by the path metaphor (i.e., Steps 1 and 5 in Table 5).[7] Although the data sets were too different to perform a sound statistical comparison, the relevance paths do appear to lead to increases in MAP over simply choosing representations randomly or based on similarity to previously "viewed" information, especially when the relevance paths were short. In the next section we describe how these findings and others listed so far in this article influence the revised design of the search interface.

3.4 Revised Interface Design

Based on the study findings described so far in this section we developed a new interface schematic. The schematic is shown in Figure 4 with the proposed relevance path marked.

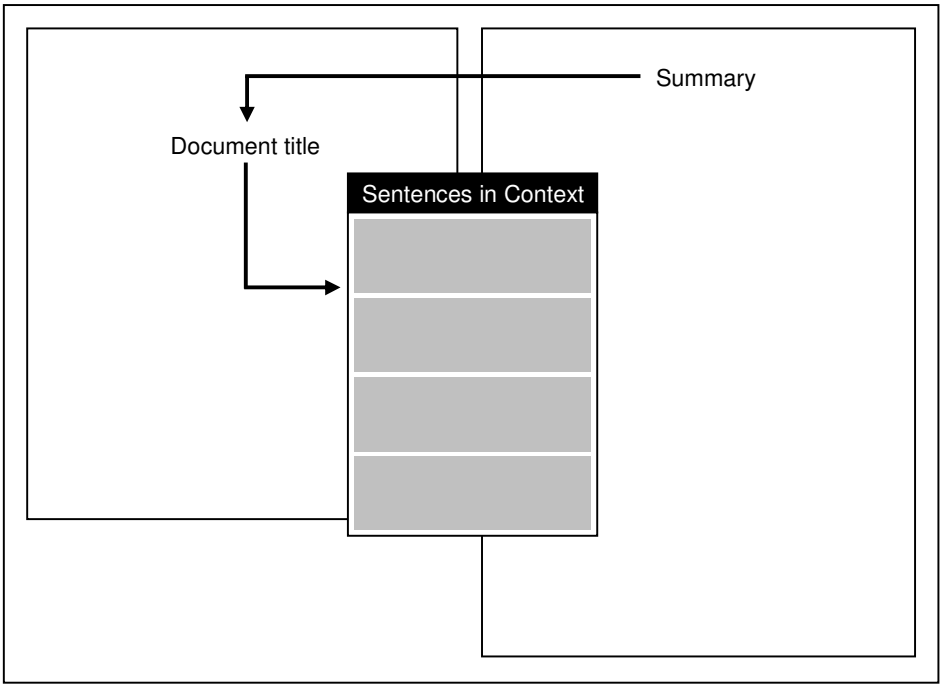

Figure 4. Revised interface schematic (with one proposed full relevance path marked).

Each of the scenarios led to the following interface design modifications:

Scenario 1: The number of representations displayed at the interface is restricted to a maximum of 50 at any given time.

[7] Since relevance paths are only between 1 and 5 representations long, it is only fair to compare Steps 1 and 5 in Table 5 with the values shown in Table 8. Comparing Steps 10 and over with Table 8 MAP values biases the comparison against relevance paths, since there is more feedback available to the representations alone.

Scenario 2: The top-ranking sentences have been removed from the initial results display and replaced by document summaries of the top-ranked documents, scored based on the terms they contain, and ranked independent of source documents. These summaries will appear on the right of the interface. Sentences in context are now shown upon hover of the document title.

Scenario 3: The arrangement of representations in the longest relevance path changes to: "Summary" *then* "Title" *then* "Sentence in Context", and the longest relevance path is shortened from five representations to three representations.

The findings of Scenarios 2 and 3 led to the removal of the top-ranking sentences and the summary sentences from the interface, since they did not add value to the IRF. In the next section we discuss our approach.

4. Discussion

In this article simulations of searcher interaction have been used as design tools to provide system designers with information about the predicted effect of design decisions.

We assessed the value of representations based on the extent to which they improved retrieval effectiveness when used for IRF. Studies on the value of document representations tend to focus on human perceptions of their usefulness for making judgements about the relevance of their source document (Barry, 1998). Different types of representations vary in length, and can hence be regarded as being more of less indicative of their sources. For example, a top-ranking sentence is less indicative than a query-based document summary (typically composed of four sentences) as it contains less information about the content of the document. The length hypothesis (Marcus *et al.*, 1978) suggests that representation quality is directly proportional to its length. However, this fails to consider the quality or nature of a representation (Janes, 1991) e.g., a document title is typically short but is assigned by an author, and may capture the key concepts in a document. The findings of our study showed that titles and summaries were the best representations of those tested as sources of RF. This is an interesting finding in itself, independent of this study and the unique nature of our interface, since many interfaces employ such representations in the presentation of search results. Titles and summaries are semantically richer representations, containing key terms, and multiple concepts or overarching themes. Top-ranking / summary sentences (and surrounding contexts) may only contain one concept that is elaborated upon in the context. Since titles and summaries are more indicative of (in this case

relevant) documents query expansion terms chosen from titles and summaries are likely to retrieve more relevant documents.

Other factors studied included the number of representations to display at the interface, and the ordering of representations in relevance paths. We found that MAP generally plateaus at a point where the current document set has been exhaustively mined as a source of query expansion terms. At this point the searcher may benefit from the generation of a new set of document representations to interact with, or the introduction of other sources of evidence such as anchor texts (Brin & Page, 1998), or overlapping retrieval results (Croft & Thompson, 1987). The overlap between the top-ranking sentences, the summary sentences, the document summaries, and sentences in context in relevance paths appears to hinder IRF effectiveness, at least with the metric that was employed in this study. Concepts that appear in these four path representations may be reinforced by searcher interaction with all of them. This may lead to an over-training problem where the IRF model focuses too much on one particular aspect of the information need, and creates queries that retrieve only documents relevant to that aspect. For this reason, shorter paths with only very limited overlap between their contents may perform better than longer paths with a greater amount of overlap.

Relevance paths and TRS were demonstrated as being useful for feedback, although perhaps not as useful as we had initially hoped. These components have been welcomed by user study participants but appear not all that effective from the IRF model's perspective. There is a trade-off between what the system wants for IRF and what the searcher wants to facilitate the exploration of information and better understand their information problem. The study described in this article has been useful for giving the IRF model a chance to influence the design of systems meant to elicit IRF. Whilst the current setup does (and should always) include a human in the design loop, it broadens the thinking of designers of such systems to consider what evidence their RF model would like. This suggests a balance in the development of information-seeking environments between system and searcher, always catering for the latter in cases of doubt, but perhaps making controversial design decisions when the predicted payoff for RF is large and guaranteed. It is also worth considering that when designing interfaces from the model's perspective we can take advantage of the computational power of machines and use additional sources of evidence that may overload human searchers.

The simulation-based approach we adopted in this article aims to emulate some aspects of searcher interaction that goes beyond the simple simulation adopted in the *ad hoc* tasks in TREC. In those tasks the searcher interaction simulated is generally only the submission of queries. Although it is possible to simulate RF in a limited way (Buckley *et al.*, 1994), these techniques are generally not suitable when more a complex interaction paradigm is adopted, or when all interaction is used as RF. The simulations presented in this article, albeit restricted, roll back the limits of what is possible in IR evaluation. All scenarios assume that searchers view information from assessed relevant documents only. It is possible to factor "wandering" behaviours into the simulation (i.e., acknowledge that searchers also look at some proportion of non-relevant information), and vary this as part of the experimental design. However, in earlier work (White *et al.*, 2005b) we showed that the inclusion of this additional behaviour does not generally affect the *relative ranking* of components; it is such relative rankings that we are interested in for the study described in this article. Other factors, such as using the amount of relevant information in documents to determine the number of representations selected from that source, are beyond the scope of this article, but make interesting avenues for future research.

Using simulations in this way calls for the development of techniques to interpret their output, and employ decision metrics to make recommendations to designers or present visualisations of simulated behaviours as in (Chi *et al.*, 2000). The current situation places a high demand on the system designer to interpret the values of the metrics used in the simulation, in this case MAP. Although other metrics such as cumulative gain (Järvelin & Kekäläinen, 2002) – or even measures of how much effort must be expended during interaction – could also be used, and perhaps combined, decisions of how to interpret the values returned by the metrics lie with the system designer. A potentially rich area of future work is the development of agents to interpret the metrics, and make design recommendations on how to maximise them. Similar research is already being conducted in the requirements engineering community (e.g., ElKoutbi *et al.*, 1999). Of course, no matter how reliable the agent may appear to be, it would still be prudent to give designers executive control over what recommendations are implemented in later iterations of the system design.

Searchers will have preferences about how the interface should be designed. Although we have ameliorated some of costs associated with including human subjects in IIR evaluation, it is worth noting that in this research we are certainly not advocating the removal of humans from the

development of IIR interfaces. The role of the human is central to all aspects of IR, in particular interface design, and that should not change.

Systems employing IRF may be able to suggest paths through the information space that will yield the best query when used for IRF. If the implicit feedback interface can predict which feedback and sequence of feedback will yield the best retrieval performance it may be prudent to recommend this to the searcher, especially in circumstances where they appear lost or overwhelmed by the available information. Whilst recommending the route through the information that is of most utility to the searcher is obviously preferred, when searcher needs are difficult to estimate then recommending a route that yields interaction that in turn leads to the most effective modified query may be a reasonable alternative. More work is required on shaping interaction design to balance the irrational desires of searchers and the strictly rational needs of RF algorithms. That is, when IRF model is unsure about what the searcher wants, then marking representations based on what the model requires to improve effectiveness may be a worthwhile alternative.

The next stage for this research is to conduct a comparative evaluation of the interface prototype described at the beginning of this article and the interface design that emerged from interpreting the output of the simulations. This will be closely coupled with experiments to validate the simulated methodology with human searchers. The use of interaction log data to adjust some of the interaction parameters in the simulation goes at least some way to replicating real searcher behaviours, but a fully fledged study designed to assess how well the simulations do this job is essential. At least two approaches for doing this are being considered: direct assessment and indirect assessment. Direct assessment would show searchers paths that the simulation predicts they may take through the available information, and ask them to assess, using instruments such as Likert scales and semantic differentials, the quality of the path predicted. Indirect assessment would involve subjects using the interface to perform their search tasks and the simulation running in the background, taking as input interaction and predicting the next step. Success would be measured based on the proportion of occasions that the simulation correctly predicts the next searcher's next move. Further research is also necessary in comparing the quality of the decisions made by IIR system designers employing simulations and those that do not. In this article the author was the designer, but to be truly useful, searcher simulations must work in a variety of experimental contexts.

5. Conclusion

In this article we have described the use of simulations of search behaviour to help system designers redesign an IRF interface implementing some aspects of polyrepresentative theory. Searcher simulations have the potential to empower designers by giving them information about the potential impact of their design decisions at a small cost. Our approach tackled three issues: how much information to show at the interface, how much of each representation type to show at the interface, and how representations should be arranged in a relevance path. It led us to a revised interface design we may not have conceived through more traditional means of requirements capture and user-centered design. A useful by-product of our approach was the ability to contrast the effectiveness of an interface-design that used relevance paths and one that did not. The results of the comparison show that relevance paths appear useful for eliciting RF but not significantly more than selecting representations randomly. The real benefit of relevance paths (and top-ranking sentences) may therefore lie in the guidance they give to searchers as they interact, and not with the feedback they elicit for use in query expansion. The interface design that results from the use of simulations in this way is one of the most rational designs that will lead to the largest improvements in retrieval effectiveness. Conducting this research has made us aware of the opportunities that simulations present and their limitations. We have described some of these in this article in the hope that they will direct researchers interested in using simulations to do so in an appropriate way.

6. References

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

Belkin, N.J., Cool, C., Croft, W.B., and Callan, J.P. (1993). The effect of multiple query representations on information retrieval system performance. In *Proceedings of the 16th Annual ACM International Conference on Research and Development in Information Retrieval*, pp. 339-346.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research, 8 (3)*. 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 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 292-300.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7): 107-117.

Callan, J. (1994). Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 302-309.

Chi, E.H., Pirolli, P., and Pitkow, J. (2000). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computer Systems*, pp. 161-168.

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 ACM SIGCHI Conference on Human Factors in Computer Systems*, pp. 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 ACM SIGCHI Conference on Human Factors in Computer Systems*, pp. 505-512.

Cleverdon, C.W., Mills, J. and Keen, E.M. (1966). *Factors determining the performance of indexing systems*. Cranfield, UK: Aslib Cranfield Research Project, College of Aeronautics (Volume 1: Design; Volume 2: Results).

Cosijn, E. and Ingwersen, P. (2000). Dimensions of relevance. *Information Processing and Management*, 36(4): 533-550.

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

ElKoutbi, M., Khriss, I., and Keller, R.K. (1999). Generating user interface prototypes from scenarios. In *Proceedings of the 4th IEEE International Symposium on Requirements Engineering*, pp. 150-157.

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.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory, *Journal of Documentation*, 52, 3-50.

Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer: Berlin.

Janes, J.W. (1991). Relevance judgements and the incremental presentation of document representations. *Information Processing and Management*, 27(6): 629-646.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4): 422-446.

Jeffrey, R. C. (1983). *The logic of decision*. Chicago: University of Chicago Press.

Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M., Secker, J. and Walker, S. (1995). Interactive thesaurus navigation: intelligence rules ok. *Journal of the American Society for Information Science and Technology*, 46(1): 52-59.

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

Kelly, D., Dollu, V.D., and Fu, X. (2005). The loquacious user: a document-independent source of terms for query expansion. In *Proceedings of the 28th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 457-464.

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*, pp. 317-325.

Larsen, B. and Ingwersen, P. (2002). The boomerang effect: retrieving scientific documents via the network of references and citations. In *Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 397-398.

Liddy E.D. (1990). Anaphora in natural language processing and information retrieval. *Information Processing and Management*, 26(1): 39-52.

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*, pp. 324-332.

Marcus, R.S., Kugel, P. and Benenfeld, A.R. (1978). Catalog information and text as indicators of relevance. *Journal of the American Society of Information Science*, 29, 15-30.

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. *Information Retrieval,* 6, 199-223.

Muller, M.J., Wildman, D.M., and White, E.A. (1993). 'Equal opportunity' PD using PICTIVE. *Communications of the ACM*, 36(4): 64-65.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Johnson, D. (2005). Terrier information retrieval platform. In *Proceedings of the 27th European Conference on Information Retrieval*, pp. 517-519.

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*, pp. 727-734.

Paice, C.D. and Jones, P.A. (1993). The identification of important concepts in highly structured technical papers. In the *Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 69-77.

Rubenstein, R. and Hersh, H. (1984). *The Human Factor: Designing Computer Systems for People*. Digital Press.

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 213-220.

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

Siegel, S. and Castellan, N. J. 1988 *Nonparametric statistics for the behavioural sciences.* Singapore: McGraw-Hill.

Tianmiyu, M.A. and Ajiferuke, I. Y. (1988). A total relevance document interaction effects model for the evaluation of information retrieval processes. *Information Processing and Management*, 24(4), 391-404.

Turtle, H. and Croft, W.B. (1990). Inference networks for document retrieval. In *Proceedings of the 12th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1-24.

Voorhees, E.M. and Harman, D.K. (2005). *TREC: Experiment and evaluation in information retrieval*. MIT Press.

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

White, R.W., Jose, J.M., and Ruthven, I. (2003b). A granular approach to web search result presentation. In *Proceedings of the 9th IFIP TC13 Conference on Human Computer Interaction*, pp. 213-220.

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

White, R.W., Ruthven, I., and Jose, J.M. (2005a). A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 57-64.

White, R.W., Ruthven, I., Jose, J.M., and Van Rijsbergen, C.J. (2005b). Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23(3), 325-361.

White, R. W., Jose, J. M., and Ruthven, I. (2005c). Using top-ranking sentences to facilitate effective information access. *Journal of the American Society for Information Science and Technology*, 56(10), 1113-1125.

White, R. W., Jose, J. M., and Ruthven, I. (2006). An implicit feedback approach for interactive information retrieval. *Information Processing and Management*, 42(1), 166-190.

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*, pp. 249-256.