Implicit Feedback for Interactive Information Retrieval

Ryen William White

Department of Computing Science
Faculty of Computing Science, Mathematics and Statistics
University of Glasgow

Submitted for Degree of Doctor of Philosophy
at the University of Glasgow
October 2004

To my parents.

Acknowledgements

My Ph.D. has been a three year journey. I have endured short periods of frustration that made me question whether it was worthwhile, but mainly happiness that made me realise of course it was. I took this journey on my own, but some people deserve a special mention for their guidance and support along the way.

Firstly, my supervisors Joemon Jose, Ian Ruthven and Keith van Rijsbergen. Joemon, thank you for being a great supervisor, for having enough faith in my abilities to give me this chance and for all your efforts on my behalf. Ian, thank you, you are an inspiration; your feedback, support and advice surpassed any requirements of a second supervisor. Keith, thank you for your wisdom, encouragement and guidance, it has been a privilege to have the benefit of your counsel.

I would like to thank members of the Information Retrieval Group, past and present, for making the trip a pleasant one. Thank you all for your friendship, support and interest in my work.

The administration and support staff in the Department of Computing Science deserve a huge mention for keeping everything running smoothly. As does my funding body, the Engineering and Physical Sciences Research Council and the Royal Society of Edinburgh for awarding me a J.M. Lessells Travel Scholarship.

To all my friends; thank you for sustaining a genuine interest in my work and for being understanding when it intruded on my social life. I hope now that it’s over, I can be the friend that you all have been to me.

Thank you to Simone Bittman, whose understanding, patience, love and affection played an integral part in my happiness during this journey. I look forward to many more happy times.

Finally, my parents William and Sarah; words cannot truly express how much I owe you both. You gave me life and have done nothing but support me throughout it. Thank you so much for your unstinting love, help and encouragement. This one’s for you guys!

Abstract

Searchers can find the construction of query statements for submission to Information Retrieval (IR) systems a problematic activity. These problems are confounded by uncertainty about the information they are searching for, or an unfamiliarity with the retrieval system being used or collection being searched. On the World Wide Web these problems are potentially more acute as searchers receive little or no training in how to search effectively. Relevance feedback (RF) techniques allow searchers to directly communicate what information is relevant and help them construct improved query statements. However, the techniques require explicit relevance assessments that intrude on searchers' primary lines of activity and as such, searchers may be unwilling to provide this feedback. Implicit feedback systems are unobtrusive and make inferences of what is relevant based on searcher interaction. They gather information to better represent searcher needs whilst minimising the burden of explicitly reformulating queries or directly providing relevance information.

In this thesis I investigate implicit feedback techniques for interactive information retrieval. The techniques proposed aim to increase the quality and quantity of searcher interaction and use this interaction to infer searcher interests. I develop search interfaces that use representations of the top-ranked retrieved documents such as sentences and summaries to encourage a deeper examination of search results and drive the information seeking process.

Implicit feedback frameworks based on heuristic and probabilistic approaches are described. These frameworks use interaction to identify needs and estimate changes in these needs during a search. The evidence gathered is used to modify search queries and make new search decisions such as re-searching the document collection or restructuring already retrieved information. The term selection models from the frameworks and elsewhere are evaluated using a simulation-based evaluation methodology that allows different search scenarios to be modelled. Findings show that the probabilistic term selection model generated the most effective search queries and learned what was relevant in the shortest time.

Different versions of an interface that implements the probabilistic framework are evaluated to test it with human subjects and investigate how much control they want over its decisions. The experiment involved 48 subjects with different skill levels and search experience. The results show that searchers are happy to delegate responsibility to RF systems for relevance assessment (through implicit feedback), but not more severe search decisions such as formulating queries or selecting retrieval strategies. Systems that help searchers make these decisions are preferred to those that act directly on their behalf or await searcher action.

Table of Contents

Table of Figures

Table of Tables

Part I

Introduction

In this part I present an introduction to the thesis and the general outline of its structure. The background and motivation for the research are then presented. I describe the query formulation process and associated problems; feedback mechanisms designed to resolve these problems; the effects of information need development, relevance and tasks on information seeking behaviour, different forms of result presentation and interactive evaluation. Where appropriate the contents of this part motivates, and is related directly to, the work presented in later parts of this thesis.

Chapter 1

Introduction and Outline

1.1 Introduction

A searcher approaches an Information Retrieval (IR) system with a need for information derived from an 'anomalous state of knowledge' (Belkin *et al.*, 1982). This need is typically transformed into a query statement, submitted to the system and a set of potentially relevant documents is retrieved and presented. The transformation of this need into a search expression, or query, is known as *query formulation*. Through such transformations and further interaction searchers can conduct Interactive IR (IIR), where they engage in dialogue with the IR system and it dynamically responds to their feedback (Borlund, 2003).

However, search queries are only an approximate, or 'compromised' information need (Taylor, 1968), and may fall short of the description necessary to retrieve relevant documents. This problem is magnified when the information need is vague (Spink *et al.*, 1998) or searchers are unfamiliar with the collection makeup and retrieval environment (Furnas *et al.*, 1987; Salton and Buckley, 1990). On the World Wide Web (the Web) searching can be even more difficult since most Web searchers receive little or no training in how to create effective queries. Consequently, search systems need to offer robust, reliable methods for query modification.

Relevance feedback (RF) (c.f. Salton and Buckley, 1990) is the main post-query method for automatically improving a system's representation of a searcher's information need. The technique assumes the underlying need is the same across all feedback iterations (Bates, 1989) and generally relies on explicit relevance assessments provided by the searcher (Belkin *et al.*, 1996b). These indications of which documents contain relevant information are used to create a revised query that is more similar to those marked and discriminates between those marked and those not. The technique has been shown to be effective in non-interactive environments (Buckley *et al.*, 1994), but the need to explicitly mark relevant documents means searchers may be unwilling to directly provide relevance information. The user interface challenge is therefore to provide an easy and effective way to control the use of RF in systems that implement it.

Implicit RF, in which an IR system obtains relevance feedback by passively monitoring search behaviour, removes the need for the searcher to explicitly indicate which documents are relevant (Morita and Shinoda, 1994; Kelly and Teevan, 2003). The technique uses implicit relevance indications, gathered unobtrusively from searcher interaction, to modify the initial query. Traditionally, 'surrogate' measures such as document reading time, scrolling and interaction have been used to provide implicit evidence of searcher interests (Claypool *et al.*, 2001; Kelly, 2004). However, such measures are context-dependent (Kelly, 2004), vary greatly between searchers and are hence difficult to correlate with relevance across searchers and searches. Whilst not being as accurate as traditional 'explicit' RF, implicit RF (or *implicit feedback*) can be an effective substitute for its explicit counterpart in interactive information seeking environments (White *et al.*, 2002b).

This thesis is an investigation of implicit feedback methods for interactive information retrieval. Unlike the surrogate methods described above, interaction with the results interface and not with the retrieved documents is used as feedback and the only assumption I make is that searchers will view information that relates to their needs; their interests can be inferred by monitoring what information they view. Information about what results are relevant is obtained implicitly, by interpreting a searcher's selection of one search result over others as an indication that result is more relevant. The Ostensive Model (Campbell and Van Rijsbergen, 1996) is based on such principles and uses passive observational evidence, interpreted by the model, to adapt to searcher interests.

In this thesis I propose novel methods of result presentation, query modification, retrieval strategy selection and evaluation. These methods aim to facilitate effective information access and assist searchers in formulating query statements and making new search decisions on how to use these queries. Although the Web is used as the document collection for this investigation the findings are potentially generalisable to different document domains.

Interface techniques are developed and tested that encourage interaction and aim to generate an increased quality and quantity of evidence for the implicit feedback methods devised. These techniques present a variety of query-relevant representations of documents such as titles, sentences and summaries that are accessible by the searcher at the results interface.

Implicit feedback *frameworks* are created that use interaction with these representations and the traversal of paths between these representations as evidence to select terms for query modification and to make decisions on how to use the revised query. This is made possible since the interface components in the search interfaces I create are smaller than the full-text of documents, allowing relevance information to be conveyed more accurately. The frameworks proposed are divided into two parts: *term selection* (i.e., the selection of important words to modify the query) and *retrieval strategy selection* (i.e., making search decisions about how to use the query).

The term selection models from the frameworks are evaluated objectively using a novel simulation-based evaluation methodology that emulates searcher interaction. The best performing model is chosen to be further tested in a user experiment with human subjects and in three RF systems that implement the same implicit feedback framework, but offer different interface support. This evaluation tests the term selection model that estimates information needs and a component to estimate changes in needs and track these changes during a search session. It also investigates task effects and how much control searchers want over three of the central search activities associated with RF systems: conveying relevance information, creating search queries and making new search decisions about using these queries (i.e., selecting retrieval strategies).

In the remainder of this chapter I provide an outline of this thesis and describe the contribution it makes to IR research.

1.2 Outline

This thesis addresses issues in the interaction between searchers and RF systems. Traditional RF systems use searcher indications of what information is relevant as evidence for their algorithms. However, since the provision of relevance assessments is adjunct to the process of seeking information it can problematic to get searchers to communicate their preferences. Search systems that gather relevance information implicitly may be a viable alternative to traditional RF. These systems can reduce or remove the burden of making many search decisions whilst retaining the iterative process of feedback that makes RF a powerful search technique.

The research presented in this thesis focuses on the development of interfaces to Web search systems such as Google, MSN Search and Yahoo! that are important information access tools for a large number of computer users. Web searchers typically receive no formal training in query formulation and can struggle to find relevant documents. It is therefore important to develop techniques to help such searchers locate relevant information.

This thesis tackles this problem through the development of search interfaces that encourage a closer examination of search results and the creation of implicit feedback frameworks to proactively support searchers. Simulated studies and studies with human subjects are conducted to test the effectiveness of components in the frameworks I propose. In this section I describe the contents of this thesis under three general headings: *interaction*, *feedback* and *evaluation*.

1.2.1 Interaction

Traditionally, search results are presented as a ranked list of documents and searchers typically exhibit limited interaction with these lists (e.g., clicking on only a few document titles). Studies have shown that increasing the amount of interaction with retrieved information can lead to more effective searching (Spink *et al.*, 1998; White *et al.*, 2003b). In this thesis novel interface techniques are proposed that aim to encourage an increased quantity and quality of interaction with search systems. The improved interaction can be used by searchers simply to find relevant information or by implicit feedback frameworks as evidence to allow them to make decisions for the searcher. I call the approach that facilitates this interaction *content-driven information seeking*.

Content-driven approaches drive searchers to the resolution of their needs by the provision of query-relevant document representations and interface support mechanisms to adapt their presentation at the results interface when presented with new relevance information. These representations are typically sentence-based and in Chapter Three I describe the method used to select the top-ranking (or best) sentences from each document. In Chapter Four, three user studies of techniques to use these sentences to support online searching and to convey system decisions are presented and the findings used to motivate research later in the thesis. In Chapter Five the content-driven approach is extended to include more document representations and I present *content-rich* search interfaces that encourage searchers to follow paths between document representations and explore search results more fully; interaction with these representations at the results interface is used as implicit feedback. In the next section I provide an outline of the techniques used to gather this feedback.

1.2.2 Feedback

Traditional RF approaches (Salton and Buckley, 1990; Belkin *et al.*, 1996b) require searchers to explicitly mark search results as relevant. This can be a burden and searchers may feel uncomfortable with the additional control (Beaulieu and Jones, 1998). Implicit feedback alleviates this problem by making inferences on what is relevant from interaction. However, traditional implicit feedback methods such as document reading time can be unreliable and context dependent (Kelly, 2004). In this thesis I propose two implicit feedback frameworks that make decisions based on the information (e.g., sentences, document titles, document summaries) searchers interact with. The frameworks estimate current information needs and changes in these needs during a search session.

The implicit feedback approaches presented in Chapter Four use interaction of searchers to generate an internal query that dynamically updates the interface. The usability of these techniques and subject comments from studies of them motivated the development of more sophisticated mechanisms for inferring searcher interests. In Chapters Six and Seven I present heuristic-based and probabilistic implicit feedback frameworks that build on the work in Chapter Four and select query terms on the searcher's behalf.

Information needs can be dynamic and may change in a dramatic or gradual way during a search session (Bruce, 1994; Robins, 1997). In such circumstances searchers may want to reorganise or recreate the information they are viewing and assessing. RF systems typically only offer searchers the choice to re-search and generate a new set of documents. This is only one way to use this relevance information and for small need changes this may be too severe; retrieval strategies that reflect the degree of change may be more appropriate. As well as creating new query statements, the frameworks employ mechanisms to identify how much the topic of the search has changed. They can use predicted extent of the change to choose retrieval strategies that may assist in finding relevant information.

The RF techniques discussed in this thesis have the potential to alleviate some of the problems inherent in explicit relevance feedback whilst preserving many of its benefits. The initial query is still modified to become attuned to searcher needs based on an iterative process of feedback. However the information on the relevance of document representations is conveyed unobtrusively and the way the new query is used depends on the extent to which the information need is predicted to have changed (i.e., search results can be reorganised as well as recreated).

The techniques proposed are evaluated with human subjects in interactive evaluations and with simulated subjects in non-interactive evaluations where appropriate. In the next section I provide an outline of how these techniques are used in this thesis.

1.2.3 Evaluation

In total, six evaluations are conducted as part of this thesis; five involving human subjects and one involving simulated subjects (i.e., user simulations that emulate searcher interaction). Human subjects are used in circumstances where I am interested in gathering qualitative data on subject opinion (via questionnaires or interviews) or quantitative data on search behaviour (via interaction logs and my observations). Simulated subjects are used when I require direct control over search strategies and want to evaluate model performance without influence from unwanted external factors. In Chapter Four and Chapter Nine, user experiments are described during which subjects provide their perceptions of the experimental systems and recommendations for future improvements. The experiments investigate: (i) the performance of the implicit feedback frameworks, and (ii) how subjects perceive and adapt to the interface components and interface support mechanisms for relevance assessment, query formulation and retrieval strategy selection. The experimental systems used are described in Chapter Ten and the results are presented and discussed in Chapters Eleven and Twelve.

In Chapter Eight I describe a simulation-based study that uses a novel evaluation methodology to assess components in the implicit feedback frameworks (and other baselines) that select query modification terms for the searcher. The approach simulates interaction with the search interface described in Chapter Five and tests how well the frameworks perform in a variety of pre-determined retrieval scenarios. In the next section I describe the overall layout of the thesis.

1.3 Overall Layout

This thesis is divided into five parts:

Part I: Introduction

This part comprises Chapters One and Two. It provides the background and motivates work described in this thesis.

Part II: Facilitating Effective Information Access

This part contains Chapters Three and Four. It begins by describing the techniques used to extract and choose the query-relevant Top-Ranking Sentences that are used in interfaces throughout this thesis (Chapter Three). This part describes the content-driven information seeking approach used to facilitate interaction with the retrieved documents, and discusses the findings of three related user studies that demonstrate its effectiveness (Chapter Four). This part also contains an overview of the search interfaces that generate evidence for the implicit feedback frameworks presented in Part III (Chapter Five).

Part III: Implicit Feedback Frameworks

In this part I describe the implicit feedback frameworks that use searcher interaction with the search interface described in Chapter Five to modify queries and make new search decisions. Two frameworks are described; one based on pre-defined heuristics (Chapter Six) and one probabilistic (Chapter Seven). A simulation-based evaluation to benchmark the term selection components of these frameworks also forms part of Part III (Chapter Eight).

Part IV: User Experiment

In this part I present a user experiment that investigates the framework whose term selection component was chosen in Chapter Eight, different forms of interface support for presenting the decisions it makes and issues of searcher control in the interaction with feedback systems implementing the framework. The hypotheses and experimental methodology are presented (Chapter Nine) and the experimental systems described (Chapter Ten). The results of the experiment (Chapter Eleven) and the discussion of them (Chapter Twelve) are also included in this part of the thesis.

Part V: Conclusion

This part comprises Chapters Thirteen and Fourteen. The conclusions drawn from the user experiment in Part IV and the thesis overall are described (Chapter Thirteen), and avenues for future work are identified (Chapter Fourteen).

Chapter 2

Background and Motivation

2.1 Introduction

This thesis is an investigation of implicit feedback methods for interactive information retrieval. Novel methods of result presentation, query modification, retrieval strategy selection and evaluation are all proposed. The interface methods described aim to facilitate effective information access and assist searchers in formulating query statements and choosing retrieval strategies such as re-searching document collections or restructuring the already retrieved information.

This chapter provides the background for the research described in this thesis and creates a context within which the work is situated. It contains sections on query formulation and associated problems; feedback mechanisms designed to resolve these problems; the effects of information need development, relevance and tasks on information seeking behaviour, interactive evaluation and different forms of result presentation. Where appropriate the content of this chapter motivates, and is related directly to, the work presented in later chapters in this thesis. This chapter begins by addressing issues in the creation of query statements for submission to retrieval systems.

2.2 Query Formulation

The value of systems that help searchers find relevant information is becoming increasingly apparent. Such systems involve a searcher, with a need for information, motivated by a gap in their current state of knowledge (Belkin *et al.*, 1982), seeking the information required to close the gap, solve the problem that initiated the seeking and satisfy their need. Typically, searchers are expected to express this need via a set of query terms submitted to the search system. This query is compared to each document in the collection, and a set of potentially relevant documents is returned. These documents may not be completely relevant, and it is the relevant (or partially relevant) parts that contribute most to satisfying information needs.

Traditional Information Retrieval (IR) systems assume a model of information seeking known as ‘specified searching’ (Oddy, 1977), where the query presented to the system is assumed to be a specification of the type of information searchers are trying to retrieve. When the searcher is unsure of how relevant documents have been indexed and stored in the IR system, retrieval can be difficult. This problem is more acute on the Web where searchers are typically untrained, unaware of what documents exist and how these documents have been indexed by commercial search engines.

The relative success of IR systems can depend on at least two factors: (i) the question posed by the searcher, and (ii) the searcher’s ability to successfully interpret the response offered. If (i) and (ii) are handled well then the probability of a successful search is increased. In reality, this scenario is often not realised. IR systems work on a ‘quality-in, quality-out’ principle (Croft and Thompson, 1987) where a query more attuned to the searcher’s real information needs will produce better results. However, searchers may be unable to adequately define the characteristics of relevant documents, or indeed any relevant information. In such cases, the searcher’s information needs are said to be *ill-defined*. The results of Wilson (1981) made the cognitive processes behind such resultant vague, uncertain and unclear searches an important theme in IR research.

A search is motivated by an incompleteness (Mackay, 1960; Taylor, 1968; Ingwersen, 1992) or ‘problematic situation’ (Belkin, 1984) in the mind of the searcher that develops into a desire for information. When a search begins a searcher’s state of knowledge is in an ‘anomalous state’, and they have a gap between what they know and what they want to know. This gap is a situation-driven phenomenon, known as their *information need*. A way of satisfying this need can be found via relevant documents and any accumulation of knowledge en route to the final answer, including the perusal of partially relevant and even irrelevant documents. The need is prone to develop or change during this time and evolves from an initial, vague state into one known and understood by the searcher (Ingwersen, 1994). As the information need evolves the searcher’s ability to articulate query statements improves based on his or her level of understanding of the problem (Belkin, 2000).

The formulation of query statements can be a cognitively demanding process resulting in queries that are approximate, or ‘compromised’ representations of information needs (Taylor, 1968). To model the creation of the search query, Taylor suggests a continuum where searchers' abilities move initially from questions, to problems, to finally sense-making, although the boundaries between these three stages appear blurred (Muller and Thiel, 1994). Kuhlthau (1999) found in an empirical study that cognitive uncertainty increases during the initial stages of a search due to interpretative problems with the retrieved data. When the information needs are vague (Spink *et al.*, 1998), there is an anomalous state of knowledge (Belkin *et al.*, 1982), or searchers are unfamiliar with the collection makeup and retrieval environment (Furnas *et al.*, 1987; Salton and Buckley, 1990) problems with query formulation are magnified.

The widespread use of commercial search systems has brought IR, and the associated problems with query formulation, to the general user populace of the World Wide Web. Search engines such as Google, Yahoo! and MSN Search have grown in popularity and process millions of queries daily. However, the users of such systems typically receive no formal training in how to create queries, exhibit limited interaction with the results of their searches and fail to use the advanced search features that many Web search engines provide (Jansen *et al.*, 2000). Silverstein *et al.* (1999) demonstrated that searchers rarely browse beyond the first page of results and submit short queries composed of a small number of query terms. The standard interaction metaphor with Web search engines is one in which searchers submit many queries and briefly examine the results obtained.

Broder (2002) proposed a taxonomy of Web searches containing different types of queries. He suggested that queries can be *navigational* (to reach a particular site), *informational* (to acquire information present on one or more sites) and *transactional* (to perform some Web-mediated activity). Commercial search engines are designed for navigational and known-item searches where the searcher may well be able to formulate queries without assistance. However, as Broder suggests, only around a quarter of searches actually fall into this category. Over half are for informational purposes, where searchers may be unable to form queries to express their knowledge lack. IR systems, especially those on the Web, where search experience of searchers may be low, should offer methods for query modification that help searchers devise a query that represents their information needs. In this thesis I introduce new techniques to help searchers do this. In the next section I describe relevance feedback, the most commonly used technique to assist in the formulation of effective query statements.

2.3 Relevance Feedback

Search systems operate using a standard retrieval model, where a searcher, with a need for information, searches for documents that will help supply this information. As described in the previous section searchers are typically expected to describe the information they require via a set of query words submitted to the search system. This query is compared to each document in the collection, and a set of potentially relevant documents is returned. It is rare that searchers will retrieve the information they seek in response to their initial retrieval formulation (Van Rijsbergen, 1986). However, such problems can be resolved by iterative, interactive techniques. The initial query can be reformulated during each iteration either explicitly by the searcher or based on searcher interaction.

The direct involvement of the searcher in interactive IR results in a dialogue between the IR system and the searcher that is potentially muddled and misdirected (Ingwersen, 1992). Searchers may lack a sufficiently developed idea of what information they seek and may be unable to conceptualise their needs into a query statement understandable by the search system. When unfamiliar with the collection of documents being searched they may have insufficient search experience to adapt their query formulation strategy (Taylor, 1968; Kuhlthau, 1988), and it is often necessary for searchers to interact with the retrieval system to clarify their query.

Relevance feedback (RF) is a technique that helps searchers improve the quality of their query statements and has been shown to be effective in non-interactive experimental environments (e.g., Salton and Buckley, 1990) and to a limited extent in IIR (Beaulieu, 1997). It allows searchers to mark documents as relevant to their needs and present this information to the IR system. The information can then be used to retrieve more documents like the relevant documents and rank documents similar to the relevant ones before other documents (Ruthven, 2001, p. 38). RF is a cyclical process: a set of documents retrieved in response to an initial query are presented to the searcher, who indicates which documents are relevant. This information is used by the system to produce a modified query which is used to retrieve a new set of documents that are presented to the searcher. This process is known as an *iteration* of RF, and repeats until the required set of documents is found.

To work effectively, RF algorithms must obtain feedback from searchers about the relevance of the retrieved search results. This feedback typically involves the explicit marking of documents as relevant. The system takes terms from the documents marked and these are used to expand the query or re-weight the existing query terms. This process is referred to as *query modification*. The process increases the score of terms that occur in relevant documents and decreases the weights of those in non-relevant documents. The terms chosen by the RF system are typically those that discriminate most between the documents marked and those that are not. The query statement that evolves can be thought of as a representation of a searcher's interests within a search session (Ruthven *et al.*, 2002a).

The classic model of IR involves the retrieval of documents in response to a query devised and submitted by the searcher. The query is a one-time static conception of the problem, where the need assumed constant for the entire search session, regardless of the information viewed. RF is an iterative process to improve a search system's representation of a static information need. That is, the need after a number of iterations is assumed to be the same as at the beginning of the search (Bates, 1989). The aim of RF is not to provide information that enables a change in the topic of the search.

The evolution of the query statement across a number of feedback iterations is best viewed as a linear process, resulting in the formulation of an improved query. Initially, this model of RF was not regarded as an interaction between searcher and system and a potential source of relevance information. However current accounts of feedback in IIR expand the notion of feedback to one in which the system and the searcher engage in direct dialogue, with feedback flowing from searcher to system and vice-versa (Spink and Losee, 1996).

The value of IIR systems that use RF over systems that do not offer RF has already been established (Koenemann and Belkin, 1996). As this study demonstrates, it is possible to gain a deeper understanding of what searchers want from RF systems through empirical investigation. A number of studies have found that searchers exhibit a desire for explicit relevance feedback features and, in particular, term suggestion features (Hancock-Beaulieu and Walker, 1992; Koenemann and Belkin, 1996; Beaulieu, 1997; Belkin *et al.*, 2000). However, evidence from these and related studies have indicated that the features of RF systems are not used in interactive searching (Beaulieu, 1997; Belkin *et al.*, 2001; Ruthven *et al.*, 2001); there appears to be an inconsistency between what searchers say they want and what they actually use when confronted with RF systems. Searchers may lack the cognitive resources to effectively manage the additional requirements of the marking documents whilst trying to complete their search task. The interface support for explicit RF can often take the form of checkboxes next to each document at the interface, allowing searchers to mark documents as relevant, or a sliding scale that allows them to indicate the *extent* to which a document is relevant (Ruthven *et al.*, 2002b). The process of indicating which information is relevant is unfamiliar to searchers, and is adjunct to the activity of locating relevant information. The feedback mechanism is not implemented as part of the routine search activity; searchers may forget to use the feature or find it too onerous (Furnas, 2002).

Despite the apparent advantages of RF there have been relatively few attempts to implement it in a full commercial environment. Aalbersberg (1992) cited two possible reasons for this trend; the high computational load necessitated by the RF algorithms and unfriendliness of the RF interface. With recent improvements in processing power, the computational expense is no longer of real concern. Although the user interface challenge remains, technological advances mean that interfaces can be constructed that make RF more easily understood by searchers (Tague and Schultz, 1988; Gauch, 1992).

RF systems suffer from a trade-off between the searcher visiting documents because the system expects them to (i.e., to gauge their relevance) and the searcher visiting documents because they genuinely want to (i.e., they are interested in their content). This problem is perhaps more acute after submission of the first query, where the searcher is required by the retrieval system to peruse and assess documents in the first page of results. The first query is merely tentative, designed to retrieve a set of documents to then be assessed.

In operational environments searchers may be unable or unwilling to visit documents to assess their relevance. Documents may be lengthy or complex, searchers may have time restrictions or the initial query may have retrieved a poor set of documents. In RF systems the searcher is only able to judge the relevance of the documents that are presented to them. If a small number of relevant documents are retrieved then the ability of the system to approximate the searcher's information need (via modified queries taken from searchers' relevance judgements) can be adversely affected. RF systems can suffer badly if the corpus consists of a large number of multi-topic or partially relevant documents. In such documents, it is more likely that the relevant parts will contain the appropriate potential query modification terms, and terms in the remainder of the document may be erroneous, irrelevant and inappropriate. However, RF systems treat documents as single entities with an inherent notion of relevance and non-relevance encompassing the whole entity, not the constituent parts. For this reason, it may be worthwhile to base relevance assessments for such documents not on the whole document, but only on the pertinent parts (Salton *et al.*, 1993; Callan, 1994; Allan, 1995). Query-biased summarisation (Tombros and Sanderson, 1998), can reveal the most relevant parts of the document (based on the query), and also remove the need to browse to documents to assess them. The summaries may allow searchers to assess documents for relevance, and give feedback, more quickly. Similar approaches have been shown to be effective in a number of studies (Strzalkowski *et al.*, 1998; Lam-Adesina and Jones, 2001; White *et al.*, 2003b) and are used in this thesis to create many representations of documents than can be assessed through traditional implicit or explicit relevance feedback.

Relevance is an 'intuitive' concept (Saracevic, 1996) of which there are many different types (Mizzaro, 1998), and as such is not easy to define or measure. Traditional RF systems use a binary notion of relevance: either a document is relevant, or it is not. This is an overly-simplified view of what is an implicitly variable and immeasurable concept. Many studies in IR have either used binary notions of relevance directly (Rees, 1967; Schamber *et al.*, 1990), or collapsed more complex scales (incorporating the 'fuzzy regions of relevance' (Spink *et al.*, 1998)) into binary scales for analysis purposes (Saracevic *et al.*, 1988; Schamber, 1991; Pao, 1993). Partial relevance, despite its usefulness (Spink *et al.*, 1998) is typically ignored in RF systems since the formulae used to select query expansion terms and re-weight existing terms use a binary notion of relevance. There is therefore a need to incorporate less concrete, more fuzzy notions of relevance into the term selection process that underlies RF (Ruthven *et al.*, 2002b).

Another potential application of RF techniques is in *negative relevance feedback*; the selection of important terms in non-relevant documents that are then de-emphasised or removed completely from the query. This approach has been shown to not detract from, and may improve, searching behaviour when used in interactive IR applications (Belkin *et al.*, 1996a; 1998). In these studies it was suggested that the technique was difficult to use, not helpful and its effectiveness was dependent on the search topic. This may be due to how negative relevance feedback was supported at the interface.

The RF features investigated in some of the studies described in this section may have been influenced by the environment in which they were evaluated (i.e., in a controlled, laboratory setting). In a study looking at different types of query expansion techniques, Dennis *et al.* (1998) found that although searchers could successfully use novel expansion techniques and could be convinced of the benefits of these techniques in a laboratory or training environment, they often stopped using these techniques in operational environments. Anick (2003) recently found in a Web-based study, that many searchers made use of a term suggestion feature to refine their query. The results suggest the potential of term suggestion features, in some types of searching environments, especially for single session interactions. The different findings in these two studies suggest that RF may be situation-dependent and that many factors other than its usefulness influence its use. In the next section techniques to help searchers use RF systems are discussed.

2.4 Interface Support for Relevance Feedback

RF is an effective technique in non-interactive experiments (Buckley *et al.*, 1994). However, only a few studies have investigated the use of RF in interactive IR (Koenemann and Belkin, 1996; Beaulieu, 1997) and have highlighted problems in the use of RF by searchers at the interface. Typically RF systems require searchers to assess a number of documents at each feedback iteration. This activity includes the viewing of documents to assess their value and the marking of documents to indicate their relevance.

There are a number of factors that can affect the use of RF in an interactive context. Relevance assessments are usually binary in nature (i.e., a document is either relevant or it is not) and no account is taken of partial relevance; where a document may not be completely relevant to the topic of the search or the searcher is uncertain about relevance. Previous studies have shown that the number of partially relevant documents in a retrieved set of documents is correlated with changes in the search topic or relevance criteria (Spink *et al.*, 1998). Potentially relevant documents are therefore useful in driving the search forward or changing the scope of the search. The techniques used to represent the document at the interface are also important for the use of RF. Janes (1991) and Barry *et al.* (1998) demonstrated in two separate investigations that the use of different document representations (e.g., title, abstract, full-text) can affect relevance assessments. The order in which relevance assessments are made can also affect searchers' feelings of satisfaction with the RF system (Tianmiyu and Ajiferuke, 1988).

Whilst RF is conceptually simple, researchers are becoming increasingly aware that it does not provide support for the search strategies and tactics used by searchers (Bates, 1990). One problem is that the underlying query modification algorithms need a lot of relevance information to operate effectively (Rocchio, 1971). The current design of explicit RF interfaces does not fit well with this requirement, and despite their simplicity, searchers have shown a reluctance to provide relevance assessments. Beaulieu and Jones (1998) suggest that increased feedback and searcher control over query operations may increase cognitive load and that more control will not necessarily improve retrieval effectiveness. In their studies, Belkin *et al.* (2001) showed that systems suggesting terms for query expansion based on explicit feedback provided to the system were useful for searchers. However, a system implementing a *pseudo-relevance feedback* technique (that assumed the top *n* documents were relevant) was better received, leading to improved search performance and searcher satisfaction. The nature of the feedback was the only difference from the traditional explicit relevance feedback system and the pseudo-relevance feedback system which removed the burden of having to interact with the search system or mark search results as relevant. The study described in Part IV of this thesis complements this work. Rather than assuming a certain number of documents are relevant, two of the three experimental systems used in the study estimate what is relevant implicitly from searcher interaction. These systems are compared against an experimental baseline, where searchers can explicitly mark items as relevant. That is, rather than *assuming* documents are relevant, the experimental systems that use implicit feedback *infer* which are relevant, from searcher interaction.

RF is typically treated as a batch process where searchers provide feedback on the relevance of a number of documents and request support in query formulation. This may not be the best approach as in interactive environments searchers assess documents individually, not as a batch. Incremental feedback (Aalbersberg, 1992) requires searchers to assess documents individually; they are asked about the relevance of a document before being shown the next document. Through this feedback process the query is iteratively modified. The method does not force searchers to use RF although it does force them to provide feedback and may hinder their abilities to make relative relevance assessments between documents (Eisenberg and Barry, 1988; Florance and Marchionini, 1995). To resolve this problem, Campbell proposed an ostensive weighting technique (Campbell, 1999) that uses browse paths between retrieved documents to implicitly infer information needs. The paths followed through such *information spaces* are affected by the interests of the searcher.

In Campbell's system, known as the *ostensive browser*, documents are represented by nodes and the route travelled between documents by search paths. Clicking on a node is assumed to be an indication of relevance and the system performs an iteration of RF using the node clicked and all objects in the path followed to reach that node. The top-ranked documents are presented at the interface and the searcher can select one of those shown, or return to a path followed previously. There is an implicit assumption that when choosing one document that this document is more relevant than the alternatives. Ostensive relevance techniques have been used to model interaction on the Web. Azman and Ounis (2004) use data-mining techniques to test ostensive relevance profiles based on searcher logs of clicked hyperlinks. In related work, Golovchinsky (1997) also used hyperlinks clicked as indications that words in the anchor text of the link were relevant.

One of the main aims of Campbell's work on ostension was to remove the need for a searcher to manipulate a query. In contrast, Belkin *et al.* (2003) try to improve search effectiveness by encouraging searchers to produce more complete initial queries by providing more space for query entry or asking searchers to more fully describe their information problem. These techniques were successful, but still depend on the searcher's ability to conceptualise their information needs, something RF tries to address.

The process of retrieving relevant information is rich and complex (Bates, 1990; Ingwersen, 1992; Belkin *et al.*, 1993). Bates (1990) suggested that there are situations where searchers may wish to control their own search and there are situations where they would like to make use of IR systems to automate parts of their search. As suggested in Beaulieu and Jones (1998) and Fowkes and Beaulieu (2000) the level of interface support can be varied based on search complexity and associated cognitive load. In the study presented in Part IV of this thesis I compare three search systems that provide searchers with varying levels of interface support. Related empirical studies (e.g., Ellis, 1989) have shown that searchers are actively interested in their search and are keen to feel in control over what information is included or excluded and why. Other interaction metaphors (such as Rodden's use of a bookshelf to represent the current search context) have also been used to help searchers use RF systems (1998).

On the Web search systems such as Excite and Google offer relevance feedback by providing searchers with the opportunity to request 'More Like This' or 'Similar Pages' and retrieve related documents. Studies by Spink and Saracevic (1997) and Jansen *et al.* (2000) have shown that relevance feedback on the Web is used around half as much as in traditional IR searches. Therefore, the design of RF techniques for the Web needs to be more carefully approached than in other document domains as the searchers who use them are typically untrained in how to use search systems that implement them.

Systems such as Kartoo [1], the Hyperindex Browser (Bruza *et al.*, 2000), Paraphrase (Anick and Tipirneni, 1999) and Prisma (Anick, 2003) have all tried to incorporate feedback and term suggestion mechanisms into interactive Web search. Vivisimo [2] uses clustering technology to recommend additional query terms. These systems assume that Web searchers are mainly concerned with maximising relevant results on the first page (Spink *et al.*, 2002) and rely on searchers to select the most appropriate terms (selected from the most relevant documents) to express their needs. These approaches typically assume top-ranked documents are relevant (i.e., use pseudo-relevance feedback) and give searchers control over which terms are added to the query. If the initial query is poorly conceived, irrelevant documents may be highly ranked, leading to erroneous term suggestions. The techniques presented in this thesis are also Web-based, yet rather than assuming a certain number of top-ranked documents are

[1] http://www.kartoo.com
[2] http://www.vivisimo.com

relevant they make inferences on the relevance of document components from searcher interaction.

Interaction with feedback systems has an associated cost in terms of time and effort expended. Reading and rating a large number of documents is a costly activity that is not always justified by the results obtained. To be truly useful, searcher-system dialogue must have a perceived benefit to the searcher since they may depend on it directly. If this benefit cannot be guaranteed then feedback approaches based on passive observational evidence may be more appropriate. That is, feedback approaches where the searcher has no pre-conceived expectations of their performance. In previous work [3] (White *et al.*, 2002b) I have examined the extent to which *implicit* feedback (where the system attempts to estimate what the searcher may be interested in) can act as a substitute for *explicit* feedback (where searchers explicitly mark documents relevant). I side-stepped the problem of getting searchers to explicitly mark documents relevant by making predictions on relevance through analysing interaction with the system and using it to improve the effectiveness of system support. In the next section I describe the more popular measures for inferring interests from passive observational evidence.

2.5 Implicit Feedback Measures

As the previous sections have demonstrated, RF systems suffer from a number of problems that make effective alternatives appealing. Implicit feedback techniques unobtrusively infer information needs based on search behaviour, and can be used to individuate system responses and build models of system users. Implicit feedback techniques have been used to retrieve, filter and recommend different types of document (e.g., Web documents, email messages, newsgroup articles) from a variety of online sources. The research described in this section is limited to the use of implicit feedback techniques for information retrieval related tasks. In Sections 2.5 and 2.6 human actors are referred to as 'users' rather than 'searchers' since implicit feedback can also be provided whilst they are involved in activities other than searching for information.

Some of the *surrogate* measures (or behaviours) that have been most extensively investigated as sources of implicit feedback include reading time, saving, printing, selecting and referencing (Morita and Shinoda, 1994; Konstan *et al.*, 1997; Joachims *et al.*, 1997; Billsus and Pazzani, 1999; Seo and Yang, 2000). The primary advantage in using implicit techniques is that they remove the cost to the searcher of providing feedback. Implicit measures are

[3] *TRSFeedback* study in Chapter Four.

generally thought to be less accurate than explicit measures (Nichols, 1997) but as described in the previous section if implemented carefully can be effective substitutes for them (White *et al.*, 2002b). Since large quantities of implicit data can be gathered at no extra cost to the searcher, they are attractive alternatives to explicit techniques. Moreover, implicit measures can be combined with explicit ratings to obtain a more accurate representation of searcher interests.

Since implicit feedback is based on searcher behaviour there can be many possible sources for implicit evidence. Nichols (1997), Oard and Kim (2001), Claypool, *et al.* (2001) and Kelly and Teevan (2003) all provide conceptual classifications of potential behavioural sources of implicit feedback.

Nichols (1997) provided the first classification of implicit feedback by categorising the actions that a searcher might be observed performing during information seeking. Nichols discusses the costs and benefits of using implicit ratings in information seeking, and categorises these ratings by the actions a searcher may perform. He suggests that limited evidence shows there is potential in implicit rating, but that there is little experimental evidence to evaluate its effectiveness. Claypool *et al.* (2001) carried out such an evaluation and showed that certain implicit indicators could be used to infer searcher interests.

Oard and Kim (2001) built on the work of Nichols by categorising implicit ratings into four main types based on the underlying intent of the observed behaviour: *examine*, *retain*, *reference* and *annotate*. ‘Examine’ is where a searcher studies a document, and examples of such behaviour are view (e.g., reading time), listen and select. ‘Retain’ is where a searcher saves a document for later use and examples include bookmark, save and print. Further examples of keeping behaviours on the Web, where information is retained for later re-use, can be found in Jones *et al.* (2001). ‘Reference’ behaviours involve users linking all or part of a document to another document and examples include reply, link and cite. ‘Annotate’ are those behaviours that the searcher engages in to intentionally add personal value to an information object, such as marking-up, rating and organising documents.

Kelly and Teevan (2003) classify much of implicit feedback research and add another behaviour category to the four already defined in this section. Their ‘Create’ category describes the behaviours typically associated with the creation of original information. These five categories only represent a sample of the possible behaviours that searchers may exhibit, but are sufficient to classify most search behaviour. Only the ‘Examine’ and ‘Retain’ categories are appropriate to categorise the behaviour of online searchers since the

'Reference', 'Annotate' and 'Create' categories all require control over the content of documents and the structure of document spaces. Searchers rarely have this control and the work reported in this thesis aims to help searchers in interactive information seeking environments. The techniques I propose reside in the 'Examine' category and infer information needs via inferences made from the information viewed. The approach uses interaction with the results interface of the search system rather than actual documents. This allows the system to control what information the searcher observes and more closely monitor their interaction.

Claypool *et al.* (2001) categorised a series of different interest indicators and propose a set of observable behaviours that can be used as implicit measures of interest. Experimental subjects were asked to browse documents in an unstructured way. The time spent on a page, mouse clicks and scrolling were all recorded automatically by the customised browser that subjects used. Subjects were asked to explicitly rate each page before leaving it and the ratings were used to evaluate the implicit measures. The researchers found a strong positive correlation between time and scrolling behaviours and the explicit ratings assigned. However, since subjects were not engaged in a search task (just asked to browse a set of interesting documents), the applicability of the findings to information seeking scenarios is uncertain.

In general, the application of implicit measures does not consider the characteristics of individual searchers. All searchers are assumed to exhibit stereotypical search behaviours around relevant information. One of the most widely used behaviours for implicit modelling is reading time (Morita and Shinoda, 1994; Konstan *et al.*, 1997; Billsus and Pazzani, 1999; Seo and Yang, 2000; White *et al.*, 2002a). This has been questioned for being too simplistic and not taking full account the influencing effects of other factors such as task, topic and user characteristics (Kelly and Belkin, 2001; 2002). In a related study Kelly and Cool (2002) found that as topic familiarity increased, reading time decreased, and proposed that as the searcher's state of knowledge increased, their search behaviour altered. Such findings suggest a role for different relevance indications at different points in the search session, based on topic familiarity. Kelly (2004) suggested that to develop models of document preference, techniques based on implicit feedback must also be able models the searcher's information seeking context and must construct models that are personal to the searcher, not general, for all searchers. Kelly also found in the same naturalistic user study that despite its popularity as an implicit feedback measure document retention is not a good indicator of document preference. Searchers may retain a document for a number of reasons, only one of which is the relevance of its content. Morita and Shinoda (1994) conducted a longitudinal study of search behaviours when reading newsgroup documents. Over a period of time, subjects were required to view newsgroup documents and explicitly rate their interest in the articles. The authors examined reading time and keeping behaviours of experimental subjects. They found a positive relationship between reading time and user interests, but none between retention and document interests. In a related study Goecks and Shavlik (2000) measured hyperlinks clicked, scrolling performed and processor cycles used to unobtrusively predict the interests of a searcher. They integrated these measures into an agent that employed a neural network and showed that it could predict user activity and build a model of their interests that could be used to search the Web on their behalf.

The development of user models (UM) offers the potential of individuating users and tracking their information seeking behaviour and evolving information needs over time. A user model is a system generated or selected description of the user that facilitates interaction between the two (Allen, 1990). [4] Through UM, the picture developed of the user should allow the system to effectively predict user responses and lead to more effective, efficient, personalised interactions.

To gather the information necessary to create a UM, a medium of knowledge elicitation is necessary. Traditionally in IR this has been done by human intermediaries (Ingwersen, 1982; Belkin, 1984; Belkin *et al.*, 1987; Spink *et al.*, 1996) who gather knowledge from searchers by asking correctly phrased appropriate questions at opportune moments during the search. Then, once the searcher's problem has been identified they suggest appropriate retrieval strategies. The implicit feedback frameworks proposed in this thesis assume the role of a human intermediary, inferring information needs and recommending retrieval strategies.

Affective User Modelling (AUM) has created user models that incorporate the emotions of computer users (Picard, 1997). Most of the research into AUM has been based on multi-modal forms of input as affective wearables (Picard, 1997), speech recognition (Ball and Breese, 1999) and facial expression recognition (Wehrle and Kaiser, 2000). The human-computer interaction community have begun using these types of behaviours to infer attention (Fendlay *et al.*, 1995), and more recently, cognitive load (Ikehara *et al.*, 2003) and emotion (Picard and Klein, 2002). It is possible that information obtained from these types of behaviour can provide useful implicit feedback for information retrieval related tasks.

Surrogate measures such as document examination and retention can vary greatly between searchers, are dependent on the information seeking context (e.g., the document domain and

[4] Although other types of user model exist (Fischer, 2000), I focus only on this type in this thesis.

task characteristics) and can be unreliable sources of evidence for implicit feedback (Kelly, 2004). In this thesis I deal with the use of implicit feedback from searcher interaction with the results interface (e.g., clicking on hyperlinks, viewing summaries). As Kelly suggests, traditional implicit feedback measures that use interaction with the full-text of documents can be unreliable and difficult to capture, and are therefore not used in this thesis. In Chapter Four I describe a study conducted as part of the investigation of content-driven information seeking. The results of the study show that reading time is correlated with the relevance of document summaries. This result was interesting and although statistically significant required an *a priori* determination of benchmark times for each experimental subject that meant the findings were insufficiently generalisable to be used as part of the implicit feedback mechanism in the frameworks described in this thesis. These were designed to operate without prior knowledge of searcher interests or preferences, which may not always be available. In the next section a brief summary is given of attentive information systems that develop user models of searchers to infer and process their long and short-term interests.

2.6 Attentive Systems

In operational environments, systems that use unobtrusive methods to infer interests are called attentive or adaptive systems. These observe the user (via their interaction), model the user (based on this interaction), and anticipate the user (based on the model they develop). Attentive *information* systems aim to support user's information needs and construct a model based on their interaction. In attentive systems, the responsibility for monitoring this interaction is usually assigned to an external *agent* or *assistant*. Examples of such agents include Lira (Balabanovic and Shoham, 1995), WebWatcher (Armstrong *et al.*, 1995), Suitor (Maglio *et al.*, 2000), Watson (Budzik and Hammond, 2000), PowerScout (Lieberman *et al.*, 2001), and Letizia (Lieberman, 1995).

Attentive systems accompany the user during their information seeking journey, and by observing search behaviour (and other behaviours in inter-modal systems) they can model user interests. Such systems can typically operate on a restricted document domain or on the Web. The methods used to capture this interest and present system suggestions differ from system to system. Letizia (Lieberman, 1995), for example, learns user's current interests and by doing a lookahead search (i.e., predicting what searchers may be interested in the future, based on inference history) can recommend nearby pages. PowerScout (Lieberman *et al.*, 2001) uses a model of user interests to construct a new complex query and search the Web for documents semantically similar to the last relevant document. WebWatcher (Armstrong *et al.*, 1995), in a similar way, accompanies users as they browse, but as well as observing, WebWatcher also acts as a *learning apprentice* (Mitchell *et al.*, 1994). Over time the system learns to acquire greater expertise for the parts of the Web that it has visited in the past, and for the topics in which previous visitors have had an interest. Suitor (Maglio *et al.*, 2000), tracks computer users through multiple channels – gaze, Web browsing, application focus – to determine their interests. Watson (Budzik and Hammond, 2000), uses contextual information, in the form of text in the active document, and uses this information to proactively retrieve documents from distributed information repositories by devising a new query.

All of these systems can be classified as *behaviour-based* interface agents (Maes, 1994; Lashkari *et al.*, 1994), that develop and enhance their knowledge of the current domain incrementally from inferences made about user interaction. Systems of this type typically adopt a strategy that lies midway between IR and *information filtering* (IF) (Sheth and Maes, 1993). In IR, a searcher actively queries a base of mostly irrelevant knowledge in the hope of extracting a small amount of relevant information. In IF, the searcher is the passive target of a stream of mostly relevant information, and the task is to remove or de-emphasise the less relevant or completely irrelevant material. Belkin and Croft (1992) present a more detailed comparison of IR and IF.

These systems work with the user's searching/browsing in a concurrent manner, finding and presenting documents to them during the search based on system inference of relevance/current interest. Lira (Balabanovic and Shoham, 1995) contrasts with such systems in two ways; it builds a model based on users' explicit ratings, and browses the Web offline to return a set of pages that match the user's interest. It is questionable whether it is strictly an attentive information system, as it does not immediately respond to change the search topic and relies on the explicit ratings users provide.

To predict what might be useful, an attentive information system must learn from a user's history of activity to improve both the relevance and timeliness of its suggestions. Attentive systems are personalised, developing and revising a user model throughout the whole search session. As the user model evolves, becoming a closer approximation to the user after each step, it should be able to recommend new documents should a significant change in need and/or user dissatisfaction be detected. Any new suggestions should be presented to users in an unobtrusive and timely way, either selecting opportune moments of prolonged inactivity or in the periphery of the current, active task. These concepts are embodied by systems with a *just-in-time* (JIT) information infrastructure, where information is brought to users just as they need it, without requiring explicit requests (Budzik and Hammond, 2000). Such systems automatically search information repositories on the user's behalf, as well as providing an explicit, query-entry interface.

Attentive information systems can be distinguished by a few main characteristics. They are capable of gathering information on user behaviour from a number of sources, even across multiple modalities. When only a single source is used, the probability of making incorrect inference of user intentions is high. In contrast, with multiple sources of evidence (e.g., many applications open concurrently) ambiguity can be removed and a more accurate user model can be constructed.

An emerging research area is in the development of systems that provide the ability to search unified indices of a user's personal information repositories. These stores contain items such as electronic mails, Web pages, documents, images, appointments and other similar files that are amassed by the users over a period of time. Systems such as *Stuff I've Seen* (Dumais *et al.*, 2003) and *MyLifeBits* (Gemmel *et al.*, 2002) attempt to help users search these files and allow them to re-use information they have already seen. This is in contrast to many of the systems described in this section, which search vast online repositories to help searchers find information they may not own or is unfamiliar to them. Systems that search personal domains have the advantage of being able to build extensive profiles of those that use them.

A number of IR researchers have attempted to create a medium of knowledge elicitation traditionally performed by human intermediaries. From this user models can be created that can be used to select retrieval strategies (Oddy, 1977; Rich, 1983; Croft and Thompson, 1987; Brajnik *et al.*, 1987; Vickery and Brooks, 1987; Belkin *et al.*, 1993). Systems of this nature have focused on characterising tasks, topic knowledge and document preferences to predict searcher responses, goals and search strategies. These systems typically make many assumptions about the search environment in which they operate and the searchers that use them.

IR systems such as THOMAS (Oddy, 1977) and Grundy (Rich, 1983) tried to infer user preferences by characterising search behaviour. Grundy assumed homogeneity in the user population and used stereotypes to personalise retrieval. Systems based on search stereotypes are flawed since a sample of searchers is typically heterogeneous; searchers typically have different needs and exhibit diverse search behaviours. To address the problems of user modelling based on stereotypical representations of users systems such as IR-NLI II (Brajnik *et al.*, 1987) and FIRE (Brajnik *et al.*, 1996) have attempted to individuate the user modelling process. Searcher histories were constructed across time to tailor retrieval. Systems like

PLEXUS (Vickery and Brooks, 1987) and I^{3}R (Croft and Thompson, 1987) used different methods to improve query formulation and select appropriate retrieval strategies. PLEXUS simulated a reference librarian and asked a series of questions to build a more reliable user model. I^{3}R used multiple retrieval techniques to form a better model of the searcher's information needs. Models were constructed in I^{3}R based on explicit relevance feedback about what terms and concepts were of interest to searchers. This system still required searchers to perform an active part in explicitly defining the model and their interests before using the system.

In this thesis I present techniques that operate without any domain knowledge and without *a priori* user models approved by the searcher. The techniques use only the original query of the searcher and their interaction with document representations extracted from the retrieved information to build a model of searcher interests. A number of factors can influence this interaction or more generally, information seeking behaviour of searchers. In the next section three of the most important are described in relation to this thesis: task, relevance and dynamic relevance.

2.7 Information Seeking Behaviour

In this section I review research in some aspects of information seeking behaviour that may influence the provision of RF and the use of systems that implement it. The main issues addressed are the role of the work task and the concept and dynamism of relevance.

2.7.1 Task

The underlying *work task* e.g., constructing an essay, is the motivational force behind information seeking. Simulated work tasks (Borlund and Ingwersen, 1997; 1998; Borlund, 2000b) allow personal assessments of what constitutes relevant material and the creation of a consistent information seeking context. Simulated work tasks are modifications of artificial goals that attempt to provide the searcher with a more robust description of the information problem (Vakkari, 2003). These types of task may be used in laboratory evaluations to provide search scenarios to assess search systems or sets of interface features (Pors, 2000; White *et al.*, 2003b).

In recent times the influence of the task in information seeking scenarios has been acknowledged and used to explain differences in relevance assessments and system use (Vakkari, 2001). The work task relates to the activity that results in the need for information

(Belkin *et al.*, 1982; Ingwersen, 1992). Several *search tasks* may stem from the original work task, each involving a series of decisions about system operation and search result assessment.

Vakkari (2003) identified two major options for modelling tasks as independent variables. The first is to use task complexity as a way to model tasks. This approach is related to how much the searcher knows about the information requirements, process and outcome of the task (Byström and Järvelin, 1995; Bell and Ruthven, 2004). The second is to use information search process models (ISPs), such as that of Kuhlthau (1993a), to analyse tasks and their impact on information seeking. This approach views tasks as a series of stages, relating specific behaviours to these specific stages and has demonstrated that both the type of information needed and searcher interaction vary according to task complexity and stage. The task classification used in the experiment in Part IV uses tasks of varying complexities to encourage different information seeking behaviours at different stages of the ISP.

The effect of task complexity on information seeking has already been studied (Vakkari, 1998; 1999). In his work Vakkari suggests that task complexity has an impact on how well searchers can perceive their information needs, and relates it to prior search knowledge, search strategies and relevance. He proposes that although it is possible to alter the factors that affect complexity, task complexity is not objective and personal factors such as topic familiarity, search experience and search knowledge can impact on searcher's assessments of it (Kelly and Cool, 2002; Vakkari, 2002). Investigations into which factors contribute to making a task more or less complex have been carried out by a number of researchers (Campbell, 1988; Byström and Järvelin, 1995; Bell and Ruthven, 2004).

Campbell (1988) described task complexity as a function of psychological states of the task performer, the interaction between the task characteristics and the abilities of the task performer and the objective attributes of the task itself, such as the number of sub-tasks or the uncertainty of the task outcome.

Byström and Järvelin (1995) proposed a task categorisation based on investigating real search behaviour in real work situations. The categorisation defines five levels of task complexity based on the *a priori* determinability of tasks; a measure of the extent to which the searcher can deduce required task inputs, processes and outputs from the initial task statement. Tasks that are increasingly complex encourage increased uncertainty about task inputs, search processes and outputs. Byström and Järvelin found through an examination of the task-based literature of a number of different research fields, two main groups of task characteristics related to complexity: characteristics related to the *a priori* determinability of tasks and characteristics related to the extent of tasks. They developed a qualitative method for task-level analysis of the effects of task complexity on information-seeking and found a relationship between task complexity and types of information needed, information channels used, and sources used.

Bell and Ruthven (2004) collapse the five category classification of Byström and Järvelin into three categories and test whether they can predicatively influence the complexity of artificial search tasks. They investigate the effects of task complexity on searcher perceptions and satisfaction with the search process. They find that it is possible to predict and manipulate search task complexity. In Part IV of this thesis a number of search interfaces are evaluated using varying degrees of task complexity based on the Bell and Ruthven methodology. The varying degrees of task complexity aim to encourage different information seeking behaviours. For example, one would expect searchers to exhibit browsing behaviour for complex search tasks, and focused, keyword searching for simple tasks (Kuhlthau, 1991).

Tasks have also been modelled as stages in the information seeking process. The model of the information search process proposed by Kuhlthau (1993a) characterised task performance into six stages, each of which differentiated and determined the type of information searched for, how it was searched for and how relevance assessments were made. Another popular model of the various types of information search processes that characterise a searcher's information seeking was proposed by Ellis (1989) who defined the following characteristics of information seeking behaviour: starting, chaining, browsing, differentiating, monitoring, extracting, verifying and ending. The work of Kuhlthau (1993a), Ellis (1989) and Marchionini (1995) has demonstrated that during a search people progress through a series of stages, adopting different strategies and exhibiting different information seeking behaviours as they move from one stage of the information seeking process to another. Movement from one stage to the next is not necessarily sequential; a searcher can cycle through several stages and/or skip others.

Research on implicit feedback has more or less ignored the affect of task. In many studies, the specific domain of the searcher's activities is limited and as is the task. For instance, Morita and Shinoda (1994) and others (Billsus and Pazzani, 1999; Miller *et al.*, 2003) considered the behaviour of users as they interacted with online news services like Netnews and Usenet. Kim, Oard, and Romanik (2000) studied behaviour in a more traditional information seeking task, finding sources for a research paper, and Cooper and Chen (2001) investigated how behaviour could be used as implicit feedback in an online library card catalogue. Studies that place no limits on the types of Internet searching activities investigated like Claypool, *et al.* (2001) and Joachims, Freitag, and Mitchell (1997), make no attempt to measure task, and instead, construe the task to be finding 'useful' or 'interesting' information. An exception to this is the study conducted by Kelly and Belkin (2004) which attempted to understand how reading behaviour changed with respect to specific task and topic. Studies on implicit feedback have not attempted to characterise information seeking tasks or stages, or conduct a systematic investigation of their impact on observable behaviours and relevance assessments; the user experiment in Part IV addresses some of these issues.

In the next section I consider another important factor affecting information seeking behaviour, relevance.

2.7.2 Relevance

In RF relevance is traditionally considered as a binary concept: a document is either relevant or it is not. This overly simplistic view is necessitated by query expansion algorithms and evaluation measures such as precision and recall (Spink *et al.*, 1998). Schamber *et al.* (1990) proposed relevance feedback as a multidimensional phenomenon when they discussed the role of situational relevance in making relevance assessments. Situational relevance is the usefulness of an information object to the current search task.

Saracevic (1996) identified five types of relevance: (i) system or algorithmic, (ii) topical, (iii) pertinence or cognitive, (iv) situational and (v) motivational. System or algorithmic relevance is objective and is the same regardless of searcher. The others are dependent on the searcher and their information seeking context. Topical relevance describes the level of searcher belief in the match between document content and their information needs. Pertinence is similar but dependent on a searcher's cognitive state. Situational relevance is the relationship between the current task, situation or problem and documents. Motivational, or 'affective' relevance, describes the relation between motivations, intentions and goals of a searcher and those of a document. To have such relevance documents must inspire positive feelings such as satisfaction, success and accomplishment.

Implicit feedback techniques make inferences from searcher behaviour as they are engaged in information seeking activities. Since the information sought relates to their current situation one can conjecture that the searcher is communicating (albeit implicitly) examples of information that is situationally relevant. Information with situational relevance has utility in relation to the searcher's current situation (Cooper, 1971; Wilson, 1973). Borlund (2000b) expresses situational relevance as the relationship between the searcher's perception of a work task situation and a retrieved document. The use of simulated work tasks and this notion of situational relevance allow for subjective relevance assessments in laboratory evaluations.

Searchers typically use many criteria when assessing the relevance of documents. In a recent study Tombros, Ruthven and Jose (2003c) identified categories of Web page features that searchers typically use when assessing relevance; text, structure, quality, non-textual and physical properties. The findings of their study showed that the various textual aspects of Web pages (general content, textual parts containing query terms and numbers, text in the title and headings of pages), are important for identifying the utility of pages to tasks. This demonstrates the value of page content over other features for relevance assessments and motivates the use of content to facilitate effective information access (Part II).

When engaged in information seeking activities searchers endeavour to view information relevant to their needs. Frameworks such as information foraging theory (Pirolli and Card, 1995) attempt to model how searchers search in information access environments. It suggests that searchers will use information access tools and view information as long as the perceived benefit gained from viewed information outweighed the costs involved. The theory assumes that the value and cost structure of information is defined in relation to the embedding task structure and changes dynamically over time (Bates, 1989; Schamber *et al.*, 1990). Search systems should be able to adapt dynamically to cater for these changes.

2.7.3 Dynamic Relevance

To operate effectively, implicit feedback systems must identify both the current search topic and when a search has changed (i.e., moved from one topic to another). During this change a searcher's perception of relevance may change over session time. Harter (1992) proposes that relevance judgements are a psychological state in which retrieved documents that stimulate changes in the searcher's cognitive state. The query is a one-time static conception of the problem that motivates the need, where the need assumed to be constant for the entire search session, regardless of the information viewed. RF is an example of an iterative process to improve a search system's representation of a static information need. That is, the need after a number of iterations is assumed to be the same as at the beginning of the search.

Much of the early work on RF assumed that searchers have static information needs; that the information for which they are searching does not change over the course of a search (Bates, 1989). Whilst this may be true in certain cases (e.g., where the information need is well-defined), evidence from a variety of studies on information seeking behaviours (Harter, 1992;

Spink *et al.*, 1998; Tang and Solomon, 1998) have shown that in most circumstances information needs should be regarded as transient, developing entities. Information needs 'develop' or 'evolve' constantly during a search on exposure to new information. Empirical investigations (e.g., Park, 1993; Bruce, 1994) have shown that searchers' cognitive viewpoints may change during information retrieval interaction, altering their relevance assessments.

In situations where the information need is vague or uncertain, information that searchers encounter is more likely to give them new ideas and consequently new directions to follow. The information need is typically not satisfied by a single final retrieved set, but by snippets of information gathered at each stage of the ever-modifying search. An example of this is *berrypicking* (Bates, 1989) where the information required to satisfy a query is culmination of the knowledge gleaned from documents examined during the search session.

The techniques discussed previously modify queries based on the documents marked or inferred relevant. The techniques used to select terms for query modification typically do not consider *when* a document was marked relevant: a document marked at the start of a search contributes as much to RF as a document marked relevant at the current iteration. Searcher's information needs can change or develop throughout the search, and documents marked relevant early in the search may not be good examples of what is currently relevant (Saracevic, 1975).

There is evidence for the dynamic aspect of relevance, which suggests that the types, and kinds of relevance judgments made can change as a searcher progresses through various problem solving stages. For instance, Spink (1996) found that at the initial stage of problem solving, people tended to judge more documents as partially relevant than fully relevant. Alternatively, Vakkari and Hakala (2000) examined students engaged in writing a research proposal and found that the portion of partially relevant documents remained constant while the portion of relevant references decreased. The research on relevance has also demonstrated that criteria used by subjects when selecting documents may change according to stage (Kuhlthau, 1993). Kuhlthau found that students used topical relevance to identify relevant documents at the beginning stages of the information search process and pertinence to identify relevant documents at later stages of the process. Campbell addressed the issue of developing information needs with his notion of *Ostensive Relevance* (Campbell and Van Rijsbergen, 1996; Campbell, 1999). The notion extends the probabilistic retrieval model and incorporates an 'ageing' component into the weighting of terms. The component adds a temporal dimension to relevance and gives a lower weight to documents marked as relevant earlier in the search.

Searchers' understanding of their information need is augmented as they encounter additional information during a search. Campbell (2000) suggested that this augmentation occurs to support or deny beliefs in various aspects of the need. That is, the searcher revises their beliefs in what information is relevant until it reaches an end point of redundancy. This redundancy may arise because the information need has been satisfied or it no longer has perceived importance to the searcher.

Kuhlthau (1991) proposed that the feelings of doubt, anxiety and frustration are natural and play their role in information seeking. The occurrence of these feelings has already been studied (Ford, 1980; Mellon, 1986), however this anxiety has usually been associated with a lack of knowledge of information sources and apparatus. Information seeking, by its very nature, causes anxiety because there is no definite positive outcome to the search (i.e., the searcher can be unsuccessful in finding what they seek). Her model of the ISP, introduced earlier, is in six stages and is based around cognitive and affective processes at various stages in the search. More specifically, the ISP is the searcher's activity of seeking meaning from information to extend their state of knowledge on a problem or topic. The process charts information seeking activity across a search session rather than at a point in time. This is similar to Ellis's (1989) model of information seeking behaviour which proposed the following characteristics: starting, chaining, browsing, differentiating, monitoring, extracting, verifying and ending. During the session the searcher's state of knowledge is dynamic rather than static; changing as the search proceeds. The steps in either process do not have to be taken sequentially and searchers can skip or repeat steps. Marchionini (1995, pp. 49-60) proposes another model of the information seeking process. In his model the information seeking process is composed of eight parallel sub-processes: recognise an information problem, define and understand the problem, choose a search system, formulate a query, execute search, examine results, extract information and reflect/iterate/stop. This model defines the activities at each stage and is perhaps more suitable for electronic environments than Ellis's model.

Choo *et al.* (1999) develop a model of information seeking on the Web that combines both browsing and searching. They suggest that much of Ellis's model is already implemented by components currently available in Web browsers. Searchers can begin from a Web site (starting), follow links to information resources (chaining), bookmark pages (differentiating), subscribe to services that provide electronic mail alerts (monitoring) and search for information within sites or information sources (extracting).

As the need moves through these stages RF systems should be able to describe the known relevant information and adapt to changes in the need as it is augmented by viewed information. In the techniques described in this thesis these requirements are met by the creation of separate need detection and need tracking components. The need detection component chooses terms for query modification and the need tracking component chooses retrieval strategies based on the estimated change in information needs. Needs can change in a gradual and dramatic manner. RF systems typically only give the option to use the modified query to retrieve a new set of documents. However, for small changes or developments in information needs, the standard RF activity of re-searching information repositories may be too severe and actions that suit the degree of change may be appropriate.

The way in which search results are presented has an impact on the information seeking behaviour of searchers. In the next section I discuss issues related to results presentation.

2.8 Results Presentation

Searchers are typically unwilling to visit individual documents to gauge relevance and base judgments on document *surrogates*, such as titles, abstracts (i.e., short textual summaries) and URLs, presented by the IR system. The work of Landow (1987), Furnas (1997) and Pirolli and Card (1995) have stressed the importance of giving searchers clues about what information to expect if they click a link. The surrogate information assists searchers in making decisions about what documents to visit.

IR systems were originally devised for the retrieval of documents from homogeneous corpora, such as newspaper collections or library index cards. Document surrogates were usually created by experts, such as librarians or professional cataloguers. However, the growth in size, dynamism and heterogeneity of these collections necessitated the development of automated indexing techniques. This led to a reduction in the quality of the surrogates created that was documented as early as the mid 1960's (Edmundson, 1964).

Presenting lists of document surrogates has remained a popular method of presenting search results. While conveniently packaging information and providing a ranking based on estimated utility, such lists can also be restrictive; they encourage searchers to read, interpret and assess documents and their surrogates *individually*. It may be the information in the document, *complemented* by the document surrogates that searchers require to close the knowledge gap that drives their seeking. The surrogates are an intermediate step between the submission of a query and the perusal of one or more documents returned in response to that query. In a previous study (White *et al.*, 2003a) I established that the indicative worth of the automatically generated abstracts created by search engines such as Google and AltaVista was questionable and that more complete representations of documents were required.

Abstracts can be the first few lines of each document or created using summarisation techniques. Research into summarisation (Tombros and Sanderson, 1998; Driori, 2003) has developed techniques to present query-biased or contextual summaries using sentences or sentence fragments with query terms highlighted. Marchionini and Shneiderman (1998) and Dumais *et al.* (2001) present summaries of document content if the searcher hovers over the hyperlink with the mouse pointer. These approaches were shown to be slower than traditional approaches as the searcher must explicitly request the additional information. In earlier work (White *et al.*, 2002a) I have shown that the viewing of such pop-up summaries can provide implicit feedback that can be effective for determining searcher interests.

The use of visualisation techniques such as TileBars (Hearst, 1995) or thumbnails (Woodruff *et al.*, 2001; Dziadosz and Chandrasekar, 2002) have tried to help searchers make better decisions by presenting the query term distributions in retrieved documents, or small image-based previews of the retrieved documents. Other representations of search results have been tested, such as LyberWorld (Hemmje, 1995), InfoCrystal (Spoerri, 1993) and BEAD (Chalmers and Chitson, 1992). These can present the searcher with an unfamiliar, usually graphical interface that imposes an increased cognitive burden and can therefore be difficult to use. Clustering approaches such as Grouper (Zamir and Etzioni, 1999) and Scatter/Gather (Cutting *et al.*, 1992) have been developed to better organise searcher results. However, clustering methods are slow and uninformative labelling can make clusters difficult to understand. Approaches that categorise documents (Chen and Dumais, 2000; Dumais *et al.*, 2001) have also been shown to be effective. More recently, interface techniques have progressively exposed searchers to more content of a document, helping them decide whether to visit documents (Zellweger *et al.*, 2000; Paek *et al.*, 2004).

In this thesis I present and evaluate an approach that encourages a deeper examination of documents at the results interface and blurs inter-document boundaries. The approach shifts the focus of interaction from document surrogates to document content, and rank this content regardless of its source. For this purpose it uses *Top-Ranking Sentences* taken from the top retrieved documents, ranked based on the query and presented in a list to the searcher. Top-

Ranking sentences aim to help searchers target potentially useful information. Potentially relevant sentences appear near the top of the list, guiding searchers towards the answer they seek or documents of interest. The sentences encourage interaction with the content of the retrieved document set. The approach is extended in later parts of the thesis to include content-rich search interfaces that use the Top-Ranking Sentences and other document representations to encourage a deeper exploration of the retrieved information. This interaction is used by the implicit feedback frameworks described in Part III.

The effectiveness of interactive search systems needs to be evaluated. In the next section I discuss issues in the evaluation of such systems and techniques.

2.9 Evaluation

As it is important to ensure that the searcher is considered in the design of interactive search systems, they are also important in their evaluation. Evaluation of the algorithms and indexing techniques that underlie these systems is traditionally based on the Cranfield model (Cleverdon, 1960) and use collections of documents, queries and pre-determined relevance assessments to determine the performance of the IR system. Initiatives such as the Text Retrieval Conference (TREC) (Harman, 1993) create test collections and recruit assessors to assign relevance assessments to documents based on the approach used in Cranfield. Their evaluation model uses precision and recall as relevance-based measures of effectiveness that typify a system-driven approach to developing and testing IR systems for empirical research in controlled environments (Spärck-Jones, 1981; Swanson, 1986). The Cranfield model retains control over experimental variables to allow conclusions to be drawn about the performance of underlying retrieval mechanisms. RF algorithms are tested using similar methods and a very simple model of searcher interaction based on the simulated assessment of the top-ranked documents (Buckley *et al.*, 1994). The approach is restrictive, does not model searcher interaction fully and makes assumptions that places limits on the cognitive and behavioural features of the environment in which IR systems operate (Belkin and Vickery, 1985). That is, it evaluates the underlying mechanics of the system but not the components with which searchers interact or the processes involved in the interaction.

The *relevance*, *cognitive* and *interactive revolutions* (Robertson and Hancock-Beaulieu, 1992) have highlighted respectively: (i) the incompleteness of queries in representing information needs, (ii) that needs reflect an anomalous state of knowledge in the mind of the searcher (Belkin, 1980), and (iii) that since IR systems have become more interactive, the evaluation of them has to include the searcher's interactive information searching and retrieval processes.

Borlund and Ingwersen (1997), Beaulieu, Robertson and Rasmussen (1996), Cosijn and Ingwersen (2000) and Borlund (2003) have advocated the development of alternative methods to evaluate interactive search systems with information needs that are personal to the experimental subject and can change during the search session. These researchers argue that relevance should be judged against the information need of its owner, not against the query statement developed to represent it

The Cranfield model may no longer be sufficient to develop a holistic view on what factors make an effective search system (Su, 1992). It does not deal with dynamic relevance but treats relevance as a static concept entirely reflected by the query statement. RF techniques were initially developed under such restricted conditions, where the feedback was given to improve the retrieval systems' approximation of the initial expressed information need (Salton and Buckley, 1990). However, whilst this may have a limited usefulness for the evaluation of RF algorithms this model is not suitable for the evaluation of RF systems that implement these algorithms, where interaction may be complex, needs may develop and change as the search proceeds and the opinions of experimental subjects are important.

The TREC Interactive Track was developed to create better methods for the evaluation of interactive IR systems (Harman, 1996). However, the methodology employed by the track was not well-suited for the evaluation of such systems since it constrained the interaction of experimental subjects and assessed interactive search systems on conditions more suitable for a non-interactive setting (Borlund, 2000b). In response to this Borlund proposes a hybrid evaluation approach that combines experimental control, the searcher, the dynamic nature of information needs and relevance assessments, as a reasonable setting for an alternative evaluation approach of IIR systems (2003). She uses measures such as Ranked Half-Life and Relative Relevance (originally proposed by Borlund and Ingwersen (1998)) as complementary measures for recall and precision for the measurement of effectiveness of IR performance. These measures allow both subjective and objective types of relevance to be incorporated in IIR evaluation. In the user experiments presented in Parts II and IV of this thesis one of Borlund's experimental components – *simulated work task situations* – are used to create search scenarios that allows different search systems and interfaces to be compared by subjects on the basis of situational relevance.

Search systems can also be tested in *longitudinal* evaluations where an information problem is assumed to persist over a period of days, weeks, months or even years. In such circumstances searchers are likely to explore a particular topic at a 'problem-level' (Robertson and Hancock-Beaulieu, 1992) beyond a single search or search session. There have been few studies of information seeking behaviour over an extended period of time (Ellis, 1989; Smithson, 1990; Kuhlthau, 1991; 1999; Kelly, 2004). Studies of this nature can be useful in investigating searcher behaviour or evaluating search systems in operational environments. However, due to a lack of control over experimental conditions they may not be suitable for comparative evaluations such as those presented in this thesis.

Experimental approaches centred on experimental subjects will always be important in the evaluation of interactive systems. However, there has been a recent trend in using searcher simulations to test the effectiveness of retrieval systems and in particular RF approaches (Magennis and van Rijsbergen, 1998; Ruthven, 2003; Mostafa *et al.*, 2003; White *et al.*, 2004b). It could be argued that the provision of relevance judgements in the Cranfield model is a crude form of searcher simulation, where simulated searchers mark certain documents as relevant and the resultant effect on precision and recall is monitored. However, simulation-based evaluation methodologies allow more complex interactions to be modelled than the standard Cranfield approach. User experimentation can be time-consuming and costly; rather than replacing human subjects, simulation-based methodologies can simulate complex interaction and retrieval scenarios and ensure that only the best or most differently performing models are evaluated using them. In Part III of this thesis I present a novel simulation-based evaluation methodology to assess the performance of implicit feedback models in different pre-determined scenarios.

2.10 Chapter Summary

In this chapter I have described the background and motivation behind the work presented in this thesis. There is a need for techniques that will help searchers search more effectively yet reduce the burden placed on them directly to reduce the number of search decisions they must make.

RF systems suffer from a number of problems that make implicit feedback an appealing alternative. The most prevalent is that it depends on a series of relevance assessments made *explicitly* by the searcher. The nature of the process is such that searchers must visit a number of documents and explicitly mark each as either relevant or non-relevant. This is a demanding and time-consuming task that places an increased cognitive burden on those involved (Morita and Shinoda, 1994).

RF is an iterative process to improve a search system's representation of a static information need. That is, the need after a number of iterations is assumed to be the same as at the beginning of the search. The aim of RF is not to provide information that enables a change in the need itself (Bates, 1989). Traditional RF systems require the searcher to instruct the system to perform RF, i.e., perform query modification and produce a new ranked list of documents. However, this is only one way of using relevance information and may not always be appropriate. Information needs are dynamic and can change in a dramatic or gradual manner (Harter, 1992; Bruce, 1994). For gradual changes, the generation of a new result set is perhaps too severe, and revisions that reflect the *degree* of change may be more suitable.

In this thesis I tackle many of the issues addressed in this chapter. Techniques are proposed to help searchers formulate their queries. Searchers do not have to explicitly assess and mark documents as relevant; these documents are not the finest level of granularity and the way the new query is used depends on the extent to which the information need is perceived to have changed (i.e., the systems do not simply re-search). Content-driven techniques are used to encourage interaction with potentially useful parts of documents that can be used as implicit feedback. I evaluate the term selection models with a simulation-based evaluation methodology and user-centred evaluations of systems that implement them. Interface support methods are tested that vary how searchers provide relevance information, formulate queries and make search decisions on query use to establish how they want search systems that use implicit feedback to communicate their decisions.

The presentation techniques proposed in this thesis use query-relevant sentences to encourage access to retrieved information. In Part II I begin by describing how these sentences are selected and how their provision at the results interface can be used to facilitate effective information access.

Part II

Facilitating Effective Information Access

So far in this thesis I have introduced information retrieval (IR), relevance feedback (RF) and implicit feedback measures for IIR. In this part an approach is proposed to facilitate searcher interaction with the retrieved documents through the use of document representations such as query-relevant *Top-Ranking Sentences* extracted from Web documents. I call this approach 'content-driven information seeking' and it tries to encourage more interaction with search results. The approach is evaluated in three related user studies, and the findings discussed. Motivated by the success of these techniques in the user studies, I also extend this work and present an overview of a search interface that uses these techniques to present these representations to searchers and allows them to follow interactive *relevance paths* between them.

Chapter 3

Top-Ranking Sentences

3.1 Introduction

Query-relevant *Top-Ranking Sentences* chosen from top-ranked retrieved search results are used as an interface component to assist searchers throughout this thesis. These sentences are selected based on the searcher's query, facilitate access to potentially relevant information and encourage a deeper examination of search results. Documents returned in response to a query by the search system are used to create the Top-Ranking Sentences. [5] These documents are downloaded and all sentences from each document are extracted. Each sentence is assigned a score, using the scoring methodology described later in this chapter. This uses factors such as position of the sentence in document, the presence of any emphasised words and any terms that occur both in the sentence and the document title. In addition sentences receive additional scores depending on the proportion of query terms they contain. This component ensures the scoring mechanism treats sentences that use query words as important.

In this chapter I describe the Top-Ranking Sentences, give the reasons why sentences, and not other semantic entities, such as paragraphs, were chosen, and provide details on how sentences were extracted and scored. It is possible to use different presentation strategies to show these sentences to the searcher; this chapter begins with a description of the strategies used.

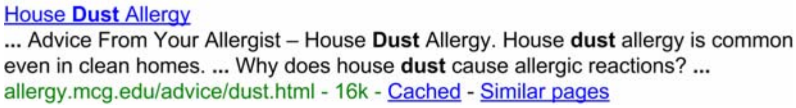
3.2 Presentation Strategies

Two presentation strategies are adopted in the interfaces described in this thesis: sentences combined to form a summary for each document and as a list across documents.

[5] The sentences selected are therefore dependent on the document ranking algorithms used by the underlying search system.

3.2.1 Sentences as Document Summary

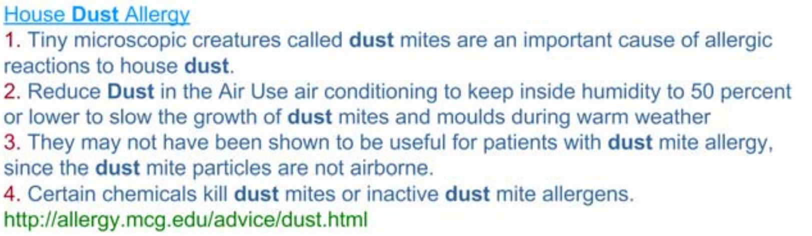
The Top-Ranking Sentences are chosen from each document and are presented at the interface for each document. The sentences combine to form a summary of the document. In response to a searcher's query, Web search engines typically only present results that consist of *document surrogate* information such as short sentence fragments and meta information similar to that shown in Figure 3.1.



House **Dust** Allergy
... Advice From Your Allergist – House **Dust** Allergy. House **dust** allergy is common even in clean homes. ... Why does house **dust** cause allergic reactions? ...
allergy.mcg.edu/advice/dust.html - 16k - Cached - Similar pages

Figure 3.1. Web search engine result for the query 'dust allergies'.

Search engines such as Google use query-biased techniques e.g., (Tombros and Sanderson, 1998) to select these sentence fragments and present query terms in the context they occur in the document. To provide this context, such systems use leading and trailing non-query terms to create short snippets of text centred on the query. These snippets, separated by ellipses, are combined to construct the document summary. This information – along with document title and the uniform resource locator (URL) – is used by searchers when deciding which documents to visit. The importance of showing searchers clues of the information resident in the source document has already been established in Landow's work on *rhetoric of departure* (Landow, 1987) and Furnas's work on *information scent* or *residue* (Furnas, 1997). Figure 3.2 shows one way in which these Top-Ranking Sentences can be used to form a summary of a retrieved document.



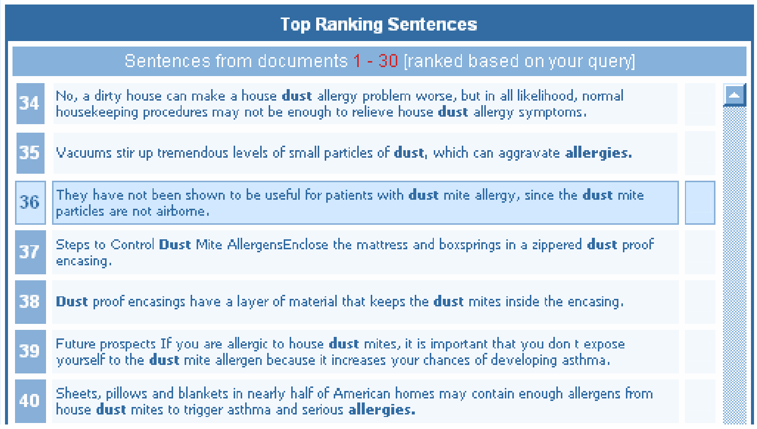
House **Dust** Allergy
1. Tiny microscopic creatures called **dust** mites are an important cause of allergic reactions to house **dust**.
2. Reduce **Dust** in the Air Use air conditioning to keep inside humidity to 50 percent or lower to slow the growth of **dust** mites and moulds during warm weather
3. They may not have been shown to be useful for patients with **dust** mite allergy, since the **dust** mite particles are not airborne.
4. Certain chemicals kill **dust** mites or inactive **dust** mite allergens.
http://allergy.mcg.edu/advice/dust.html

Figure 3.2. Sentences as document summary for the query 'dust allergies'.

In an earlier user study I demonstrated that using the best four Top-Ranking Sentences as a Web document summary was preferred to the presentation strategies exemplified in Figure 3.1 (White *et al.*, 2003b). In this user study, I found that the increased information allowed searchers to make more reliable relevance assessments, experience more satisfying searches and search more effectively. This presentation strategy groups sentences based on their source document. However, it is also possible to present Top-Ranking Sentences in a ranked list, independent of source document. In the next section I describe this approach.

3.2.2 Sentences as List

Presenting Top-Ranking Sentences independent of source documents allows highly relevant sentences from lower ranking documents, which may never be viewed simply because of their resident document's rank position, to be made accessible to the searcher. Figure 3.3 shows part of a list of Top-Ranking Sentences taken from one of the three user studies described in Chapter Four.



Top Ranking Sentences

Sentences from documents 1 - 30 [ranked based on your query]

34 No, a dirty house can make a house **dust** allergy problem worse, but in all likelihood, normal housekeeping procedures may not be enough to relieve house **dust** allergy symptoms.

35 Vacuums stir up tremendous levels of small particles of **dust**, which can aggravate **allergies.**

36 They have not been shown to be useful for patients with **dust** mite allergy, since the **dust** mite particles are not airborne.

37 Steps to Control **Dust** Mite AllergensEnclose the mattress and boxsprings in a zippered **dust** proof encasing.

38 **Dust** proof encasings have a layer of material that keeps the **dust** mites inside the encasing.

39 Future prospects If you are allergic to house **dust** mites, it is important that you don t expose yourself to the **dust** mite allergen because it increases your chances of developing asthma.

40 Sheets, pillows and blankets in nearly half of American homes may contain enough allergens from house **dust** mites to trigger asthma and serious **allergies.**

Figure 3.3. A portion of a list of Top-Ranking Sentences for the query 'dust allergies'.

The sentences are numbered based on their rank position and shown individually in the list, with query terms highlighted.

Presenting sentences in this way provides a high level of granularity, removing the restriction of document boundaries and shifting the focus from the document as a semantic entity to the information the document contains. This means that searchers are not forced to access information through documents but through the actual content of documents. Through ranking this information with respect to the query, the searcher is given an overview of the content of the returned set. A document list is biased towards the searcher's information need at the document level; documents that are a close match to the searcher's query appear near the top of the list. Presenting lists of Top-Ranking Sentences biases at the sentence level; sentences that are a close match to the searcher's query are shown near the top of a ranked list of sentences. As will be described in Chapter Four the sentences can also be used to facilitate access to low-ranked documents and communicate the effects of relevance feedback decisions.

In this thesis both sentence presentation strategies are used to assist searchers. In the next section I explain why sentences were chosen as an interface component.

3.3 Why Sentences?

Earlier studies have shown that using semantically richer document representations can be beneficial to searchers and allow them to make more reliable relevance assessments (Spink *et al.*, 1998; White *et al.*, 2003b). In this thesis sentences are used as a component to construct representations of documents that encourage searchers to examine search results more closely.

The rationale behind sentence extraction is to find a subset of the source document that represents its contents or the query, typically by scoring words and then sentences according to specific rules. The rules mainly concern the identification of clues for the importance of each sentence in the source document. Sentence extraction methods are capable of producing acceptable summaries that are domain independent (Luhn, 1958; Edmundson, 1969; Rush *et al.*, 1971; Paice, 1981; Brandow *et al.*, 1995; Salton *et al.*, 1997). This makes them perhaps more suitable for heterogeneous collections such as the Web than language generation (McKeown *et al.*, 1995) or artificial intelligence (Tait, 1985) techniques that display only a marginal level of usefulness within their restricted domains.

Research on automatic sentence extraction is well-documented. In the approach described in this chapter, sentences were used as interface components for two reasons: (i) they are by definition a coherent linguistic entity to overcome problems with semantics and present the query terms in context, (ii) they are small enough to allow searchers to assess relevance in a short time. These are preferred to paragraphs (as used in passage retrieval (Salton *et al.*, 1993; Callan, 1994)) simply because they take less time to assess. This allows searchers to make speedy judgements on the relevance/irrelevance of the information presented to them. Sentences are also the preferred semantic entity for analysis and retrieval in linguistic-based IR (Smeaton, 1990) and in the Novelty Track at the Text Retrieval Conference (TREC) (Harman, 2002).

Sentences are also used in multi-document summarisation approaches, where sentences pooled from a number of documents are used to provide a summary of these documents. Such summaries are relatively short, use domain-specific methods to score sentences (Radev and McKeown, 1998) and place a strong emphasis on coherence (Goldstein *et al.*, 2000). Sentences can also be used to form summaries of Web document clusters, as one application of the methods described in this chapter suggests (Osdin *et al.*, 2002).

In the next section I describe how sentences were selected by the search system.

3.4 Selecting Sentences

To form a list of Top-Ranking Sentences I use a sentence extraction model similar to that proposed by Tombros and Sanderson (1998). The approach extracts sentences from the top-ranked Web documents retrieved in response to a searcher's submitted query. The Web was used as searchers had experience interacting with Web documents, effective baseline search systems were readily available and realistic search scenarios for user evaluations could be easily created.

This section describes the sentence selection architecture and the techniques used to extract and score candidate sentences. Figure 3.4 shows a general overview of the approach used.

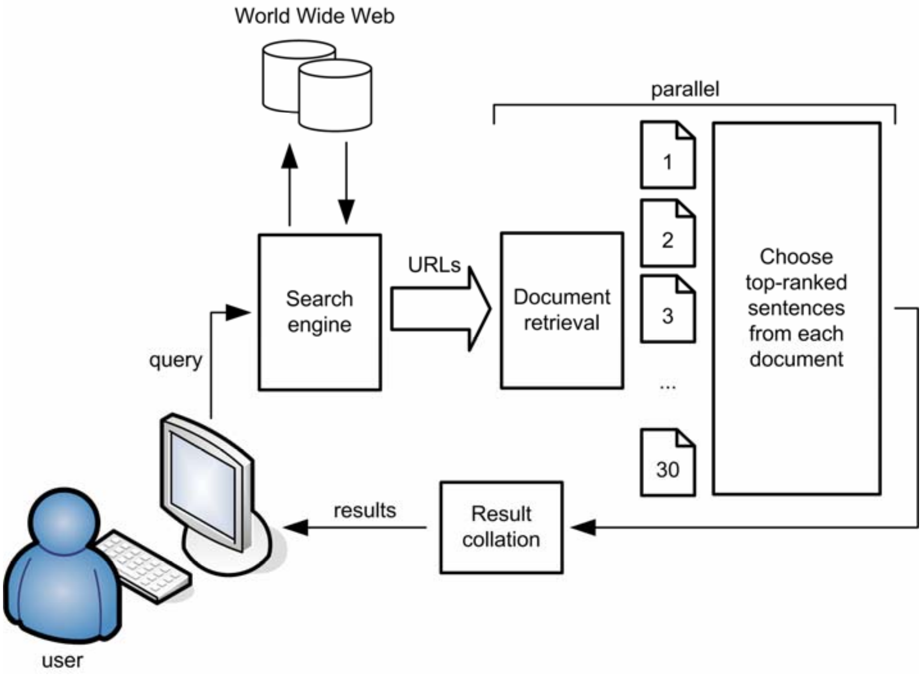


Figure 3.4. Top-Ranking sentence selection architecture.

A searcher's query statement is first passed to a Web search engine, which returns a set of documents. The documents are then visited by the system in parallel and the resident sentences extracted. The sentences are scored according to how useful they will be in reflecting page content and relevance assessment. Sentence extraction has been shown to have useful applications in Web document summarisation (Berger and Mittal, 2000). Extraction mechanisms are useful for selecting the potentially useful parts of Web documents as they can handle small portions of information and are domain independent. The extraction methods used standard punctuation (e.g., full stop, exclamation mark and question mark) and first character capitalisation methods to determine where sentences start and stop.

3.4.1 Sentence Scoring

Sentences are scored based on four criteria; *title* (e.g., sentence terms that co-occur with the title), *location* (e.g., where a sentence resides in a document), *relation to query* (e.g., the proportion of query words a sentence contains) and *text formatting* (e.g., the additional formatting added by the document author). Each scoring method is now described.

3.4.1.1 The Title Method

This method assumes the author of a document reveals the main concepts in the title of their work. It also assumes that when an author divides his work into sections, he does so in a standard manner, selecting appropriate headings for each of these divisions. Sentences containing terms that appear in the title and headings are given more weight than those without. Edmundson (1969) experimented with this method using a collection of technical documents, and assigned a greater importance to terms that appear in the title than in the section headings. The final sentence score for each sentence could then be found through the sum of the weights of each title word in the sentence. It was thought reasonable to use this method to score the sentences in Web documents as the document author has control over the title of the document and the content of the page. The title may not provide enough information on its own or supplemented with other meta-information (as in traditional result lists) to be truly indicative, but it may contain some important keywords.

3.4.1.2 The Location Method

This method assumes that: (i) that sentences located under certain headings in a document convey significant content and are therefore relevant, (ii) that important sentences tend to occur near the start, or near to the end, of a document and its paragraphs (Edmundson, 1969; Brandow *et al.*, 1995). This method assigns positive weights to words occurring under headings in a document (represented by the `<H1>`...`<H6>` HTML [6] tags) and computes the *heading weight*. As well as this, the method also assigns weights to sentences based on their ordinal position in the document (the *ordinal weight*), i.e., the first and last paragraphs in the

[6] HyperText Markup Language (HTML).

document and the first and last sentences in the paragraphs. Paragraph termination is detected in Web documents using instances of the `</P>` and `<BR><BR>` HTML tags. The total location method score for a sentence is the combination of the heading and ordinal weights.

3.4.1.3 The Text Formatting Method

The rationale behind this method stems from the idea that a Web document author may emphasise important terms (or keywords) in some way. When using the HTML that most Web documents are written in, the author can format text in a number of ways, such as **bolded**, *italicised* and underlined.

When formatted terms occur in a sentence, the sentence score is incremented by a small amount for each term. The values used were chosen based on beliefs about the value of this evidence and through pilot testing. If a term is formatted in two or more ways, say bold and italic, then the score for that sentence is incremented for each piece of formatting separately.

3.4.1.4 The Query-Biased Method

This method assumes that if searchers could see the sentences in which their query terms appeared they would be able to make a better assessment of document relevance. Tombros and Sanderson (1998) proposed a method for calculating a query score for each sentence in the document, based upon its relevance to the query. The larger the number of query terms in a sentence, the more relevant the sentence is likely to be.

The top scoring sentences are selected until the desired number of sentences is reached. This is defined to be 15-20% of the document length, or a maximum of four sentences and concurs with previous work (Edmundson, 1964; Brandow *et al.*, 1995; Kupiec *et al.*, 1995).

A potential drawback of using query-biased approaches to summarise *documents* is the biased view of the document that results; only those sentences containing many query terms are promoted. The resultant effect is a representation of the document that may not be indicative of the actual document and the emphasis therein. This problem is made more acute if the documents contain information on a variety of topics, one of which happens to be the topic of the need. Paice (1990) refers to this as the ‘coverage and balance’ problem, and is a flaw of the extracting approach. Also, it is possible that sentences containing the query terms can be scattered throughout the document. Document summaries composed of these sentences may have no cohesion and simply represent as much of the text as possible (Amitay and Paris, 2000).

3.4.1.5 Summary of Methods So Far

So far in this chapter I have described four heuristic-based methods to score the sentences extracted from Web documents. I conducted a pilot test to evaluate the sentences chosen by this approach and combined the best Top-Ranking Sentences from each document to form a document summary. Joining the sentences in this way is only one possible use of Top-Ranking Sentences and other applications are described in later chapters of this thesis. Summaries were presented to subjects as part of an interface to the Google [7] and AltaVista [8] search systems and compared with traditional forms of result presentation, where lists of titles, sentence fragments and URLs (similar to Figure 3.1) were presented. Subjects found the enriched summaries useful and that it encouraged them to interact with their search results more closely (White *et al.*, 2003b). However, the pilot study also revealed some minor problems, namely:

i. *Some sentences were too short.* Some highly scoring sentences were often headings that had been incorrectly labelled by the document author (i.e., not inside the appropriate tags). These sentences were too short to be indicative.
ii. *Some sentences were redundant.* The four Top-Ranking Sentences from each document were often too similar, query terms were shown in similar contexts and the value of the summary generated was diminished.

As a result, I incorporated two more methods to improve the quality of the sentences selected. These are *sentence length cut-off* and *redundancy checking*.

3.4.1.6 Sentence Length Cut-off

This method addressed problems with selecting sentences that were too short. All sentences used by the scoring methods need to be of a certain length (threshold: 15 tokens including punctuation). This is a frequently used threshold for removing captions, titles and headings (Kupiec *et al.*, 1995; Teufel and Moens, 1997). These headings are handled separately in the approach described in this chapter (see Section 3.4.1.2).

3.4.1.7 Redundancy Checking

To address problems with sentence redundancy a means of redundancy checking was used when selecting Top-Ranking Sentences. Through combining query-biased methods and techniques for reducing the level of redundancy it may be possible to select sentences that are

[7] http://www.google.com
[8] http://www.av.com

query-relevant and show the query terms in different contexts, one of which may be useful for the searcher. This can help ensure that sentences are selected in relation to the query that can also provide an overview of retrieved information.

The redundancy checking techniques used are based on those of Gong and Lui (2001). Unlike their work I do not use term frequency vectors for each document and compute the similarity to the document's vector. Since the approach does not create a generic document summary, there is no need to compute the similarity to the document. However, the approach does compute the degree of similarity to the query. The technique used is illustrated in Figure 3.5.

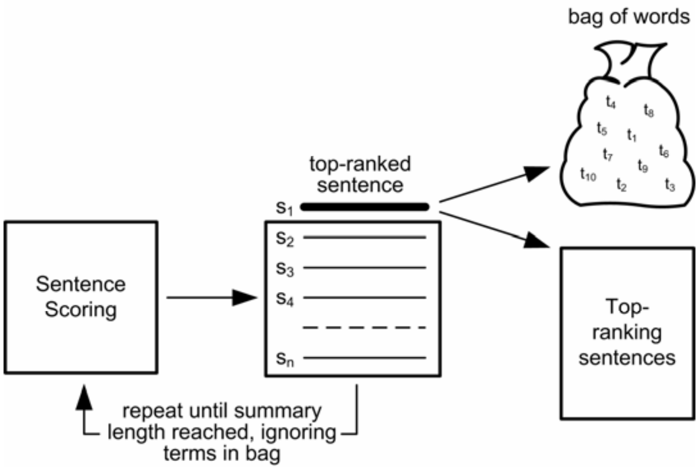


Figure 3.5. Redundancy checking in sentence selection.

The sentences extracted from the Web documents are scored based on the initial searcher query and all other methods described so far in this chapter. The sentences are then ranked based on these scores and the top sentence is removed and stored as a 'top-ranking sentence'. The *non-query* words from this sentence are placed in a bag and the process repeats, i.e., all sentences (except the one that was removed) are rescored and reordered using all constituent words that are not in the bag. The sentences chosen by this method are those that represent the query terms in different document contexts. This makes the sentences chosen suitable for document content overview (when grouped per document) or result set overview (when grouped across all top-ranked documents).

3.4.2 Combining Sentence Scores

The methods above are applied to a sentence in the sequence shown in Figure 3.6. This results in a final sentence score. The final sentence score is computed by summing together all scores from all methods. The inclusion of this scoring method had no detrimental effect on the overall sentence score should a title word *not* occur in a sentence, but a benefit if it does. All methods are given an opportunity to weight sentences; in reality a large proportion of a sentence's score is derived from its relation to the query. The redundancy checking uses all sentence scoring methods but operates independently of them and is therefore not included in the figure. The sentence length cut-off acts as a filter prior to any scoring to aid system efficiency, since only sentences of sufficient length will eventually be scored.

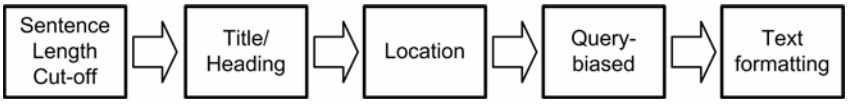


Figure 3.6. Sentence scoring methodology.

A drawback of applying a linear combination of the methods identified above is the implication that the clues provide independent items of evidence that simply needs to be combined. This may not be true, as it may be possible for the clues to interact in some way. For example, a term that is bold, underlined and in the title of the document should perhaps contribute more to its residing sentence's score than the sum of the scores for the title-keyword and twice for the text formatting (bold and underline). Despite this drawback, many studies (Edmundson, 1969; Kupiec *et al.*, 1995; Tombros and Sanderson, 1998; White *et al.*, 2003b) have used this cumulative technique to good effect for selecting sentences. In the approach presented in this thesis the chosen sentences can be used to create summaries of documents and other document representations, and presented in a ranked list, independent of source document.

3.4.3 Error Handling

The top-ranking sentence selection architecture illustrated in Figure 3.4 may experience problems selecting sentences from Web documents. This could be for a number of reasons; the document contains HTML frames, contains little or no text, or takes too long to download. [9] If this happens, or if a document is one of the restricted document types [10] then the top-ranking sentence selection architecture tries to choose sentences from the search engine's cached version of the page. The strategy employed if this is unsuccessful is dependent on the presentation strategy. In the 'Sentences as document summary' approach,

[9] The top-ranking sentence selection system rejects a Web document if it takes more than 3 seconds to download.

[10] For technical reasons, the techniques cannot select Top-Ranking Sentences from proprietary non-text files e.g., Microsoft Word documents (.doc), Microsoft Excel spreadsheets (.xls), PostScript files (.ps) and Adobe Portable Document Format files (.pdf).

the small collection of sentence fragments taken from the search engine (such as that shown in Figure 3.1) is used as a pre-created alternative to that created by the system. In contrast, in the 'Sentences as list' approach, the sentence fragments from the search engine are treated as a single sentence and included in the list of Top-Ranking Sentences as an additional entry.

3.4.4 Other Sentence Selection Methods

It is worth noting that other methods exist for selecting sentences extracted from documents. The *keyword method* (Luhn, 1958) assumes that high-frequency words that are not common *stop words* (e.g., 'of', 'the', 'and') are indicative of the document's content and are therefore useful for scoring sentences. Rather than assigning a weight to each term according to the number of times it occurs, as in (Rath, 1961; Earl, 1970), the method involves locating clusters of significant words within sentences and assigning scores to them accordingly. The query-biased approach is a version of the keyword method. Instead of providing a list of candidate index terms for each document that refer to the central concepts of the document, the searcher provides the retrieval system with a list that reflects the central concepts of the information need as they perceive it. This way, the sentences obtained from each document are those with a high score in relation to the searcher's expressed information need and have a high likelihood of relevance. The use of *syntactic criteria* (Earl, 1970), the *cue method* (Edmundson, 1969; Rush *et al.*, 1971) and the *indicator-phrase method* (Paice, 1990) rely on detailed knowledge of the corpus's language constructs and are therefore not appropriate for the heterogeneity of the Web. Paice (1990) and Spärck-Jones and Endres-Niggermeyer (1995) provide a thorough review of previous work in automatic sentence selection.

3.5 Chapter Summary

In this chapter I have introduced Top-Ranking Sentences as an interface component to present search results and encourage access to retrieved information. The rationale behind using sentences has been given, as have the techniques used to score sentences. Top-Ranking sentences can be used as document summaries, to provide an overview of the result set content and assist searchers in locating useful information. In Chapter Four I describe three user studies that use these sentences as a replacement for document lists, to communicate the effects of relevance feedback decisions and to facilitate access with retrieved documents.

Chapter 4

Content-Driven Information Seeking

4.1 Introduction

In this chapter I describe an approach that uses the techniques introduced in the previous chapter to encourage a deeper examination of the contents of the document set retrieved in response to a query. The approach shifts the focus of perusal and interaction from potentially uninformative document surrogates (such as titles, sentence fragments and URLs) to actual document content, and uses this content to drive the information seeking process. Traditional search interfaces assume searchers examine results document-by-document. In contrast the approach proposed extracts, ranks and presents the *contents* of the top-ranked document *set*. *Top-Ranking Sentences* (TRS) extracted from top documents at retrieval time are used as fine-grained representations of document content and, when combined in a ranked list, an overview of these documents. In some of the systems described in this chapter, the interaction of the searcher provides implicit relevance feedback that is used to reorder the sentences where appropriate. This chapter serves as an introduction to the use of implicit feedback in this thesis and to the style of interfaces I create.

Three related user studies with 58 different subjects were carried out to test the effectiveness of using TRS to assist searchers and communicate relevance feedback decisions. The findings of these studies were important since they influence the design of systems described in later chapters. In the analysis of the findings I focus on the relationship between the studies and qualitative subject perceptions of the approaches I describe. Hereafter I refer to the three studies as *TRSPresentation*, *TRSFeedback* and *TRSDocument*.[11] Due to variations in subjects, systems and search tasks it is difficult to make comparisons between the quantitative results obtained in each study. For this reason, quantitative results of the experiments are not

[11] *TRSPresentation* (Top-Ranking Sentences for result presentation), *TRSFeedback* (Top-Ranking Sentences for feedback decisions) and *TRSDocument* (Top-Ranking Sentences for document access).

presented in this chapter, only the subject perceptions of the techniques employed. The quantitative findings for all three studies can be found in White *et al.* (2003a) (*TRSPresentation*), White *et al.* (2002b) (*TRSFeedback*) and White *et al.* (2002a) (*TRSDocument*). This chapter describes how subjects use top-ranking sentence interfaces for their search, how this differs from traditional search methods and reason why top-ranking sentence interfaces are preferred over traditional forms of result presentation. The findings of these studies motivate the research presented in the remainder of this thesis. In the next section I describe two contrasting information seeking strategies for interacting with search interfaces; one encouraged by traditional search systems and another by systems that implement aspects of the content-driven paradigm I propose.

4.2 Information Seeking Strategies

Searchers approach IR systems with a need for information. The information required to satisfy this need transcends document boundaries and is a culmination of the knowledge gleaned from documents examined during the search session (Belkin, 1984). However, returning a ranked list of documents does not fit well with this model. The list restricts the interaction and general information seeking behaviour of searchers; they are forced to examine search results individually.

Most Web search interfaces present the searcher with little information with which to decide whether or not to view a retrieved document. Typically the only information shown is the document title, URL and short (1-2 line) sentence fragments. These fragments normally contain at least one instance of the query terms and give the searcher an idea of the context in which the query terms are used in the document.

In result lists searchers assess document relevance externally, based on what they can infer from their surrogates. On the Web, authors assign document titles and the extent to which these titles are indicative of content can vary. This differs from the static homogeneous collections used in initiatives such as TREC (Voorhees and Harman, 2000), where there is consistency in the titles/headlines assigned. Figure 3.1 (in Chapter Three) showed an example of surrogate information used in search engine result lists. This information is important since searchers use it to make decisions about what documents to view (Furnas, 1997). To provide searchers with representations that are truly indicative, it is necessary to go deeper into the documents, extracting their content at a fine level of granularity but with increased contextual coherence (i.e., with whole sentences). Through presenting full sentences to the searcher, IR systems can present the query terms in the local context in which they are used within retrieved pages.

Studies have shown that searchers refrain from using the advanced search facilities that many Web search systems offer and display limited interaction with search engine interfaces (Jansen *et al.*, 2000; Crouch *et al.*, 2002). The approach described in this chapter encourages more interaction with search interfaces and in some cases uses this interaction to make decisions on the searcher's behalf. I call this approach *content-driven information seeking* (CDIS) and it is in contrast to searcher-driven approaches where there is more onus on searchers to proactively seek information. In this section I introduce the concepts of *pull* and *push* information seeking; the latter encourages CDIS whereas the former does not.

4.2.1 Pull and Push Information Seeking

In this section two contrasting information seeking strategies are described: pull and push. The pull approach presents the searcher with surrogate document representations (e.g., titles, sentence fragments and URLs) and relies thereafter on the searcher to visit the document. In contrast, the push approach presents, and dynamically restructures, relevant content at the results interface, irrespective of source document. These strategies are affected by result presentation techniques that encourage different information seeking strategies and different emphasis. The 'need' in online searching is typically one for information. The perusal of ranked lists of documents may be an unnecessary step between query submission and direct access to this information. In what follows I describe these information seeking strategies, and the differences between them.

4.2.1.1 Pull Approach

In the pull approach the searcher must be proactive. They assess the value of documents externally based on document surrogates such as titles, sentence fragments and URLs; this requires a document-by-document examination of search results. The document is considered as the finest level of granularity and the system presents a ranked list of documents based on the estimated utility of each in relation to the searcher's submitted query.

The sentence fragments may provide the motivation with which to visit a document, however once inside the document the searcher has to locate the information then gauge its relevance in the context. Saracevic (1975) proposed, that as searchers move through the various stages of their information need evolution, where their need potentially becomes more certain (Ingwersen, 1994), their judgements of relevance are likely to change to take into account their newly encountered knowledge. Documents that are relevant at the start of the search may not be at the close. They are potentially cumbersome entities that can be completely, partially or not relevant. It may not be prudent for a searcher to spend much time reading a document to assess whether the document is relevant, and it may simply not be possible to assess a document's relevance in a short time.

In the pull approach the searcher is responsible for formulating the initial query *and* for further revising this query as the search proceeds. They are burdened with the responsibility to select additional query words and drive their own search. As suggested in Chapter Two, this can be problematic if the information need is vague (Spink *et al.*, 1998) or searchers are unfamiliar with the collection being searched or the retrieval environment (Salton and Buckley, 1990). The pull strategy is adopted by traditional search systems that, after the initial retrieval, require searchers to locate relevant information. In the next section I describe the contrasting push information seeking strategy.

4.2.1.2 Push Approach

In the push approach, the search system acts proactively, presents information extracted from the retrieved documents at query-time and restructures this information based on inferred searcher interests. Two methods are used as enabling techniques for the push paradigm; Top-Ranking Sentences and implicit feedback. In this section I describe each of these.

4.2.1.2.1 Top-Ranking Sentences

Searchers can use the Top-Ranking Sentences, selected as described in the previous chapter, to guide them through their search. The Top-Ranking Sentences provide searchers with a query-relevant overview of retrieved documents. The focus of perusal and interaction is no longer a ranked list of document surrogates offering an external view of documents. Searcher attention is instead focused on potentially useful parts of retrieved documents, meaning less time need be spent locating useful information, and more time can be spent assessing its value. These sentences can also be reordered using evidence gathered via implicit feedback from the searcher; in the next section I describe this process.

4.2.1.2.2 Implicit Feedback

As well as using the Top-Ranking Sentences to convey potentially relevant information, the sentences can also be reordered to communicate changes in the search system's formulation of relevance. Implicit feedback systems make inferences of what is relevant based on searcher interaction and do not intrude on the searcher's primary line of activity i.e., satisfying their information needs (Furnas, 2002). In traditional relevance feedback systems, the function of making judgements is intentional, and specifically for the purpose of helping the system build up a richer body of evidence on what information is relevant. However, the ultimate goal of information seeking is to satisfy an information need, not to rate documents. Systems that use implicit feedback to model information needs and enhance search queries fit better with this goal.

As already mentioned in Chapter Two, implicit feedback systems typically use measures such as document reading time, scrolling and interaction to make decisions on what information is relevant (Claypool *et al.*, 2001). However, these systems typically assume that searchers will view and interact with relevant documents more than non-relevant documents. These assumptions are context-dependent and vary greatly between searchers. The approach used for implicit feedback in this chapter makes a potentially more robust assumption: searchers will try to view relevant information. Through monitoring the information searchers interact with search systems can approximate their interests. This is made possible since the interface components the search interfaces present are smaller than the full-text of documents, allowing relevance information to be communicated more accurately.

In *TRSFeedback* and *TRSDocument* some of the experimental systems use evidence gathered via implicit feedback to restructure the retrieved information during the search. In these systems, each retrieved document has an associated summary composed of the best four Top-Ranking Sentences that appear on the interface at the searcher's request. The viewing of this summary is regarded as an indication of interest in the information it contains and is used as an indication of relevance.

These relevance indications are used by the systems to reorder the Top-Ranking Sentences. Sentences are small and the differences in sentence scores between sentences are also small. Should there be a slight change in the system's formulation of the information need a list of sentences is much more likely to change than, say, a list of documents. At no point, in any experimental system, is the searcher shown the expanded query; they are only shown the *effect* of the query (i.e., the reordered top-ranking sentence list). Reordering the sentence list based on implicit feedback means it represents the system's current estimation of the searcher's interests. Since this formulation is based solely on the viewed information the system is able to form reasonable approximations on what information is relevant. As the searcher becomes more sure of their need, or indeed as the need changes, the search system can adapt, select new query terms and use this query to update the ordering of the Top-Ranking Sentences list to reflect this change.

The user studies described in this chapter present subjects with search interfaces that may be unfamiliar to them. During these studies I felt that it was not necessary for subjects to see the contents of the modified query to use these interfaces effectively. This was the case, but some experimental subjects suggested that they may feel more comfortable with using the interfaces if they could view and manipulate the revised query. In the next section I compare the push and pull information seeking strategies.

4.2.2 Comparison of Information Seeking Strategies

The push approach extracts and presents potentially useful information to the searcher at the results interface. This content discourages searchers from examining documents individually and encourages the assessment of information resident in the result set regardless of its resident document. In contrast, the pull approach encourages searchers to assess documents externally, basing relevance assessments on the information presented in result lists.

In the push approach, sentences from documents are extracted in real-time and shown to the searcher at the results interface. In contrast, the pull approach provides less information to the searcher and they see only an external view of the document. To find relevant information, they must first visit, then locate information inside documents. The differences between the approaches are mainly in the nature of search activity and how information is presented at the search interface. Table 4.1 shows other differences.

Table 4.1
Differences between the push and pull information seeking approaches.

Factor	Approach	
	Push	Pull
Information extraction	System	Searcher
Finest granularity	Sentence	Document
Results perusal	Sentence/Scanning sentences	Document-by-document
Facilitates interaction	Sentence (content)	Surrogate
Assess document relevance	Internally	Externally
System formulation of information needs	Static/Dynamic	Static

As Table 4.1 shows, the push approach uses smaller document representations, allows searchers to assess the value of information from within documents and adapts its formulation of information needs dynamically, without searcher instruction. It is only in push systems that do not use implicit feedback techniques where the system's internal queries are static until the next searcher-initiated query iteration. The push approach selects and presents potentially relevant sentences at the results interface; visiting documents a secondary activity and the required information may be found directly at the results interface. In the pull approach, visiting documents is the main search activity and unless the task is trivial, searchers will have to visit documents to find relevant information.

In the next section I describe a series of related user studies that test the worth of the content-driven information seeking approach using Top-Ranking Sentences. These preliminary studies show that these techniques can be effective and are liked by searchers. The findings of the studies influence the design of search interfaces described later in this thesis.

4.3 User Studies

Three user studies tested the worth of Top-Ranking Sentences in different information seeking contexts. The results from these studies are summarised in this chapter, each of which utilises these sentences in a different way. In the *TRSPresentation* study the ranked sentences are used as an alternative to document lists, shifting searcher attention from the document surrogates to the document content. *TRSFeedback* uses the sentences to reflect the use of two contrasting relevance feedback techniques. Finally, *TRSDocument* uses the sentences to encourage interaction with the retrieved set, to reflect change in searcher interests and to complement, rather than replace, document lists. Each study involved real searchers and different types of information seeking scenario. The experimental systems selected Top-Ranking Sentences in real-time, when the query was submitted. This had the potential to cause delays in system operation. In each study Top-Ranking Sentences were taken from the top 30 documents to ensure the systems responded in a timely manner. In this section the generic experimental methodology is described, as are the experimental interfaces used, the tasks assigned and the relationship between studies.

4.3.1 Experimental Methodology

In all three studies human subjects were recruited from a variety of backgrounds and assigned realistic search scenarios. The length of the experiment varied between 60-90 minutes depending on the number of experimental systems. The studies followed a common experimental procedure:

i. introductory orientation;
ii. pre-search/demographic questionnaire;
iii. for each system in the study:
    a. short training session;

   b. distribute search scenario and give subjects an opportunity to clarify any ambiguities;

   c. 10-15 minutes allowed for subject to attempt the task;

   d. a post-search questionnaire;

iv. a final questionnaire, and;

v. an informal discussion (optional). [12]

There were minor differences in the methodology employed between studies, necessitated by the different experimental hypotheses.

4.3.2 Subjects

The recruitment of experimental subjects in these studies was specifically aimed at targeting two groups of subjects; *inexperienced* and *experienced*. Two out of the three studies (*TRSPresentation* and *TRSDocument*) classified subjects in this way. In these studies the classification was made based on subjects' responses on questions about their experience and their own opinion of their skill level. *TRSFeedback* did not classify subjects. The number of subjects varied between 16 and 24, the majority of whom were university students. All studies use a within-subjects experimental design meaning that subjects used all experimental systems. Latin and Graeco-Latin squares (Tague-Sutcliffe, 1992) are used to control subjects' learning effects between systems.

4.3.3 Tasks

In *TRSPresentation* and *TRSDocument* subjects attempted combinations of tasks from the following categories: *fact* search (e.g., finding a named person's current email address), *decision* search (e.g., choosing the *best* impressionist art museum) and *background* search (e.g., finding information on dust allergies) (White *et al.*, 2002a). The tasks used are included in Appendix E. Each search task was placed within a simulated work task situation, (Borlund, 2000b) that created realistic search scenarios and allowed personal assessments of what information was relevant. *TRSFeedback* was carried out as part of the TREC 2001 Interactive Track (Hersh and Over, 2001). The tasks were assigned by the track and divided up into four categories; *medical*, *buying*, *travel* and *project*. Subjects attempted a task from each category.

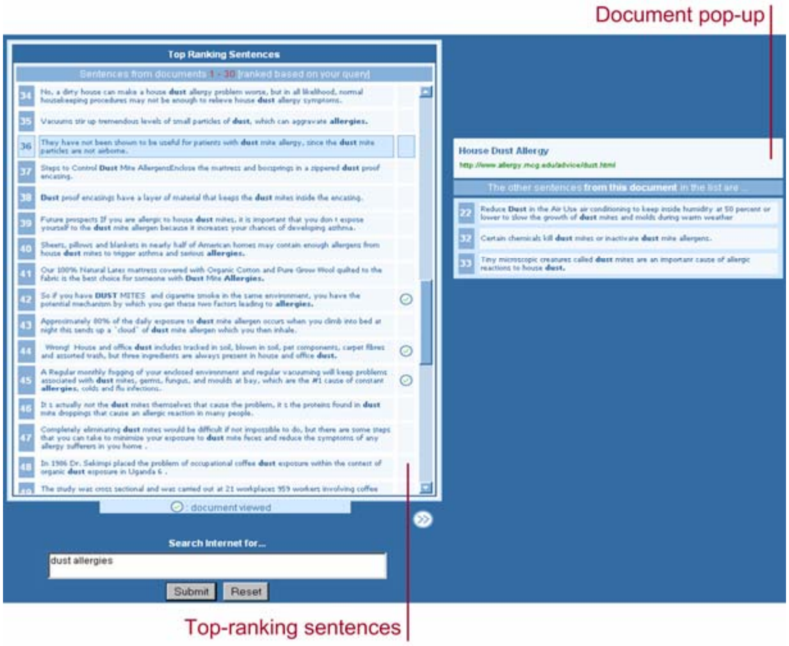
[12] The informal discussion was initiated at the subject's or experimenter's request. An opportunity to take part in such a discussion was offered to all participants.

4.3.4 Interfaces

Each of the three studies used Top-Ranking Sentences to facilitate information access, encourage interaction and convey system decisions. In this section I describe the interfaces used in each study and explain the role of the Top-Ranking Sentences in each interface. In general, the techniques described in Chapter Three are used to extract and score Top-Ranking Sentences.

4.3.4.1 TRSPresentation Study

This study investigated the effectiveness of presenting a ranked list of Top-Ranking Sentences rather than a ranked list of documents. The Top-Ranking Sentences approach is compared against two interfaces that use traditional result presentation techniques (i.e., a ranked list of document titles, summaries and URLs). One experimental system ($S_{Baseline}$) directly presents the results from the underlying search engine and the other ($S_{TRSAbstract}$) uses the Top-Ranking Sentences as a document summary, presented below the document title in the same way as in Figure 3.2 (in Chapter Three). These two systems were compared against an experimental interface ($S_{TRSList}$). This interface, shown in Figure 4.1, consists of two main components: the Top-Ranking Sentences (that replace the traditional ranked document list) and a document pop-up window, which shows the subject more information about a particular document.



Top Ranking Sentences

Sentences from documents 1 - 30 (ranked based on your query)

34 No, a dirty house can make a house **dust** allergy problem worse, but in all likelihood, normal housekeeping procedures may not be enough to relieve house **dust** allergy symptoms.

35 Vacuums stir up tremendous levels of small particles of **dust**, which can aggravate **allergies.**

36 They have not been shown to be useful for patients with **dust** mite allergy, since the **dust** mite particles are not airborne.

37 Steps to Control **Dust** Mite AllergensEnclose the mattress and boxsprings in a zippered **dust** proof encasing.

38 **Dust** proof encasings have a layer of material that keeps the **dust** mites inside the encasing.

39 Future prospects If you are allergic to house **dust** mites, it is important that you don t expose yourself to the **dust** mite allergen because it increases your chances of developing asthma.

40 Sheets, pillows and blankets in nearly half of American homes may contain enough allergens from house **dust** mites to trigger asthma and serious **allergies.**

41 Our 100% Natural Latex mattress covered with Organic Cotton and Pure Grow Wool quilted to the fabric is the best choice for someone with **Dust** Mite **Allergies.**

42 So if you have **DUST** MITES and cigarette smoke in the same environment, you have the potential mechanism by which you get these two factors leading to **allergies.** ✓

43 Approximately 80% of the daily exposure to **dust** mite allergen occurs when you climb into bed at night this sends up a "cloud" of **dust** mite allergen which you then inhale.

44 Wrong! House and office **dust** includes tracked in soil, blown in soil, pet components, carpet fibres and assorted trash, but three ingredients are always present in house and office **dust.** ✓

45 A Regular monthly fogging of your enclosed environment and regular vacuuming will keep problems associated with **dust** mites, germs, fungus, and moulds at bay, which are the #1 cause of constant **allergies**, colds and flu infections. ✓

46 It s actually not the **dust** mites themselves that cause the problem, it s the proteins found in **dust** mite droppings that cause an allergic reaction in many people.

47 Completely eliminating **dust** mites would be difficult if not impossible to do, but there are some steps that you can take to minimize your exposure to **dust** mite feces and reduce the symptoms of any allergy sufferers in you home .

48 In 1986 Dr. Sekimpi placed the problem of occupational coffee **dust** exposure within the context of organic **dust** exposure in Uganda 6 .

49 The study was cross sectional and was carried out at 21 workplaces 959 workers involving coffee

✓ : document viewed

Search Internet for...

dust allergies

Submit Reset

House Dust Allergy

http://www.allergy.mcg.edu/advice/dust.html

The other sentences **from this document** in the list are ...

22 Reduce **Dust** in the Air Use air conditioning to keep inside humidity at 50 percent or lower to slow the growth of **dust** mites and molds during warm weather

32 Certain chemicals kill **dust** mites or inactivate **dust** mite allergens.

33 Tiny microscopic creatures called **dust** mites are an important cause of allergic reactions to house **dust.**

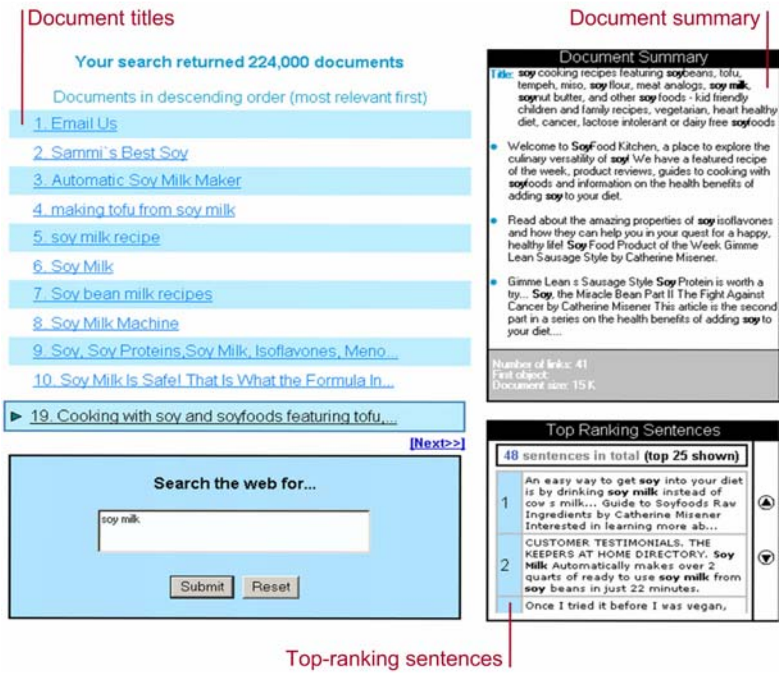
Figure 4.1. The experimental interface for the *TRSPresentation* study ($S_{TRSList}$).

The sentences are extracted and ranked using the techniques described in Chapter Three and presented in a list at the results interface. Initially there is no direct association between a

Top-Ranking Sentence in the list and its source document, i.e., there is no indication to the searcher of which document supplied each sentence. To view the association, the searcher must move the mouse pointer over a sentence. When this occurs, the sentence is highlighted and a window pops up next to it. Displaying this window next to the sentence, instead of in a fixed position on the screen, makes the sentence-document relationship more clear. In the window the searcher is shown the document title, URL and the rank position and content of any other sentences from that document that occur in the list of Top-Ranking Sentences. If no other sentences appear an appropriate message is shown. To visit a document the searcher must click the highlighted sentence, or any sentences in the pop-up window. In the $S_{TRSList}$ interface the Top-Ranking Sentences drive searcher interaction whereas in the $S_{TRSAbstract}$ and $S_{Baseline}$ systems it is the titles, abstracts and URLs that encourage searchers to interact.

4.3.4.2 TRSFeedback Study

In this study the sentences are used to communicate the effects of relevance feedback decisions. For this purpose I developed two interfaces, one where the system endeavours to estimate relevance by mining searcher interaction ($S_{Implicit}$) and one where searchers had to explicitly mark information as relevant ($S_{Explicit}$). Unlike the $S_{TRSList}$ interface described in the previous study the order of the Top-Ranking Sentences in these experimental systems updates in the presence of relevance information. The two systems adapt to the context of the search by selecting additional query terms on the searcher's behalf based on relevance information provided during the examination of results. The only difference between the two systems is in how relevance information is conveyed. The $S_{Implicit}$ system makes the assumption that the viewing of a document summary (by moving the mouse pointer over its source document title) is an indication of searcher interest in the content of the summary. The $S_{Explicit}$ system requires searchers to explicitly indicate which results are relevant by clicking on checkboxes next to each document title. Figure 4.2 shows the interface to the $S_{Implicit}$ system.



Your search returned 224,000 documents

Documents in descending order (most relevant first)

1. Email Us
2. Sammi`s Best Soy
3. Automatic Soy Milk Maker
4. making tofu from soy milk
5. soy milk recipe
6. Soy Milk
7. Soy bean milk recipes
8. Soy Milk Machine
9. Soy, Soy Proteins,Soy Milk, Isoflavones, Meno...
10. Soy Milk Is Safe! That Is What the Formula In...

▶ 19. Cooking with soy and soyfoods featuring tofu,...

Search the web for...

soy milk

Document Summary

Title: **soy** cooking recipes featuring **soy**beans, tofu, tempeh, miso, **soy** flour, meat analogs, **soy milk**, **soy**nut butter, and other **soy** foods - kid friendly children and family recipes, vegetarian, heart healthy diet, cancer, lactose intolerant or dairy free **soy**foods

- Welcome to **Soy**Food Kitchen, a place to explore the culinary versatility of **soy**! We have a featured recipe of the week, product reviews, guides to cooking with **soy**foods and information on the health benefits of adding **soy** to your diet.
- Read about the amazing properties of **soy** isoflavones and how they can help you in your quest for a happy, healthy life! **Soy** Food Product of the Week Gimme Lean Sausage Style by Catherine Misener.
- Gimme Lean s Sausage Style **Soy** Protein is worth a try... **Soy**, the Miracle Bean Part II The Fight Against Cancer by Catherine Misener This article is the second part in a series on the health benefits of adding **soy** to your diet....

Number of links: 41
First object:
Document size: 15 K

Top Ranking Sentences

48 sentences in total **(top 25 shown)**

1 An easy way to get **soy** into your diet is by drinking **soy milk** instead of cow s milk... Guide to Soyfoods Raw Ingredients by Catherine Misener Interested in learning more ab...

2 CUSTOMER TESTIMONIALS. THE KEEPERS AT HOME DIRECTORY. **Soy Milk** Automatically makes over 2 quarts of ready to use **soy milk** from **soy** beans in just 22 minutes.

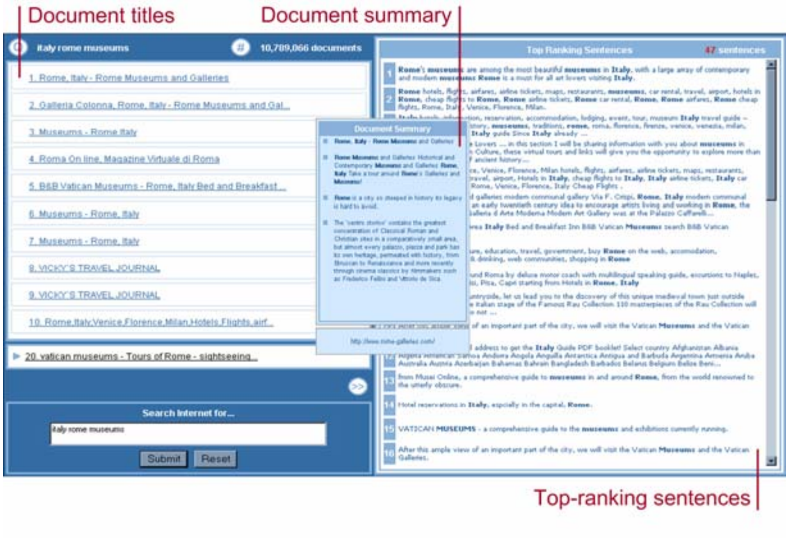
Once I tried it before I was vegan,

Figure 4.2. Experimental interface for the *TRSFeedback* study ($S_{Implicit}$).

After each relevance indication the summaries from the assessed relevant documents ($S_{Explicit}$) or assumed relevant documents ($S_{Implicit}$) are used to generate a ranked list of potential query modification terms using the *wpq* formula (Robertson, 1990). The top-ranked modification terms are chosen from this list and added to the searcher's original query. These terms are chosen from all assumed relevant summaries (i.e., those viewed so far or those from documents they have checked), and used to reorder the list of Top-Ranking Sentences based on term occurrence. The list of sentences is reordered after each relevance indication and due to the size of the window in which the sentences are displayed (shown in the bottom right-hand corner of Figure 4.2) only the top 25 sentences are displayed at any time. To make changes in the ordering of the list of sentences more noticeable, sentences from assessed summaries are removed from the list as the search progresses. The sentence reordering or the removal of Top-Ranking Sentences from the sentence list cannot be reversed by searchers.

4.3.4.3 TRSDocument Study

In a similar way to *TRSFeedback*, the experimental interface in this study uses implicit feedback techniques to gather relevance information and reorder a list of Top-Ranking Sentences. However, rather than communicating relevance feedback decisions the sentences (and the reordering) were used to facilitate access to retrieved documents. The experimental system ($S_{Feedback}$) automatically creates new search queries based on implicit feedback and is compared with a baseline summarisation system ($S_{Summarisation}$) used in White *et al.* (2003b) and a system where the order of the sentence list is static and the query is assumed to be constant within an individual search iteration (S_{Static}). Figure 4.3 shows the interface used in the S_{Static} and $S_{Feedback}$ systems. The $S_{Summarisation}$ system uses the same interface without a list of Top-Ranking Sentences.



1. Rome, Italy - Rome Museums and Galleries
2. Galleria Colonna, Rome, Italy - Rome Museums and Gal...
3. Museums - Rome Italy
4. Roma On line, Magazine Virtuale di Roma
5. B&B Vatican Museums - Rome, Italy Bed and Breakfast...
6. Museums - Rome, Italy
7. Museums - Rome, Italy
8. VICKY'S TRAVEL JOURNAL
9. VICKY'S TRAVEL JOURNAL
10. Rome,Italy,Venice,Florence,Milan,Hotels,Flights,airf...
20. vatican museums - Tours of Rome - sightseeing...

italy rome museums

1 Rome's museums are among the most beautiful museums in Italy, with a large array of contemporary and modern museums Rome is a must for all art lovers visiting Italy.

2 Rome hotels, flights, airfares, airline tickets, maps, restaurants, museums, car rental, travel, airport, hotels in Rome, cheap flights to Rome, Rome airline tickets, Rome car rental, Rome, Rome airfares, Rome cheap flights, Rome, Italy, Venice, Florence, Milan.

13 from Musei Online, a comprehensive guide to museums in and around Rome, from the world renowned to the utterly obscure.

14 Hotel reservations in Italy, espcially in the capital, Rome.

15 VATICAN MUSEUMS - a comprehensive guide to the museums and exhibitions currently running.

16 After this ample view of an important part of the city, we will visit the Vatican Museums and the Vatican Galleries.

Figure 4.3. The experimental interface for the *TRSDocument* study (S_{Static} and $S_{Feedback}$).

As in the $S_{Implicit}$ system in *TRSFeedback*, the implicit feedback in this study is the evidence the searcher gives by viewing a document summary. To allow the system to better monitor this activity, the summary was moved to a pop-up window that appears when the mouse pointer hovers over a document title and disappears when it is removed from it. Once again the *wpq* method uses this evidence to select query modification terms on receipt of this relevance information. The ordering of the sentence list changes immediately when this information is provided and coincides with the presentation of the pop-up summary window. In $S_{Feedback}$ – as in the systems in *TRSFeedback* – sentences from relevant summaries are removed from the list to make the reordering more obvious and there is no option to reverse system decisions.

In *TRSFeedback* the system interprets every summary view as an indication of relevance. This led to problems of accidental 'mouseover', with searchers passing over document titles en route to those that interested them. In this study, the system implemented a timing mechanism that dealt with this problem and allowed me to base the implicit feedback on the length of time a searcher spent viewing a summary. Subjects conducted a timing task before they used each system, allowing the calculation of a relative viewing time for each subject and the $S_{Implicit}$ system to individuate its responses. This time was used for each subject as a determinant of whether a summary they viewed was relevant. From an analysis of all subjects' viewing times from the timing task I found that they generally view relevant summaries for longer than non-relevant summaries (White *et al.*, 2002a). I use the viewing of document summaries as relevance indications since the system can easily detect which summaries are viewed and for how long.

4.3.4.4 Summary of Interfaces

All interfaces presented in this section encourage a deeper examination of search results and some use implicit feedback techniques to adapt the display in light of searcher interaction. In Table 4.2 I summarise the features of the systems created for each of the three user studies in three categories: *presentation* (i.e., how search results are presented) *summarisation* (i.e., how documents are summarised) and *feedback* (i.e., how relevance information is communicated).

Table 4.2
Features of experimental systems in the three user studies.

Feature	TRSPresentation			TRSFeedback		TRSDocument		
	$S_{Baseline}$	$S_{TRSAbstract}$	$S_{TRSList}$	$S_{Explicit}$	$S_{Implicit}$	$S_{Sum.}$	S_{Static}	$S_{Feed.}$
Presentation Method								
1. Top-Ranking Sentences			●	●	●		●	●
2. Ranked document list	●	●		●	●	●	●	●
Summarisation Method								
1. Chapter Three		●	●[α]	●	●	●	●	●
2. Search engine	●							
Feedback Method								
1. Explicit				●				
2. Implicit					●			●

[α] Although the $S_{TRSList}$ system does not present document summaries it uses the summarisation method described in Chapter Three to select Top-Ranking Sentences.

In this section we have described the experimental interfaces used in each of the three user studies. The systems within each study differ in ways necessary to test the experimental hypotheses. In the next section I describe the relationship between the three studies.

4.3.5 Inter-study Relationship

The studies all used Top-Ranking Sentences, but for a different purpose and to test different sets of hypotheses. Table 4.3 illustrates the main factors of each study.

Table 4.3
The main experimental factors in the three user studies.

	Study		
Factor	TRSPresentation	TRSFeedback	TRSDocument
Hypotheses	1. Top-Ranking Sentences are a viable alternative to Web document abstracts. 2. Top-Ranking Sentences increases awareness of result set content and is preferred by searchers. 3. Top-Ranking Sentences improve perceptions of task success, actual task success across all tasks.	1. Implicit relevance feedback is a viable substitute for explicit relevance feedback in Web retrieval.	1. The use of Top-Ranking Sentences encourages subjects to interact more fully with the retrieval results (i.e., documents) lead to more effective searching. 2. Implicit feedback improves subjects' perceptions of the system and leads to more effective interaction.
Factors measured	Search effectiveness, subject perceptions	Search effectiveness, subject perceptions	Search effectiveness, subject perceptions
Number of Systems	3	2	3
Systems (type)	1. Search engine baseline 2. TRS as abstracts 3. TRS as list	1. Implicit feedback 2. Explicit feedback	1. Summarisation baseline 2. Summarisation/TRS 3. Summarisation/TRS/ Implicit Feedback
Subjects	18	16	24
Grouping	9 inexperienced 9 experienced	None	12 inexperienced 12 experienced

Age	Average = 23.80 yrs
	Range = 32 yrs (17:49)
		Average = 24.75 yrs
		Range = 11 yrs
			Average = 24.73 yrs
			Range = 33 yrs (16:49)

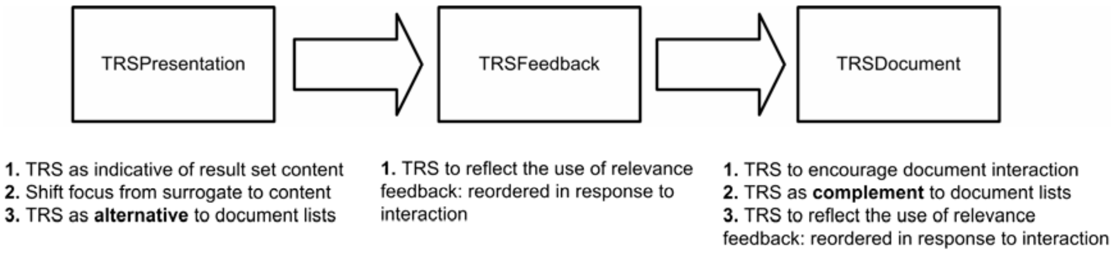
Internet Usage/week	Inexperienced = 4.2 hrs
	Experienced = 32.6 hrs
		14 hrs	Inexperienced = 4.1 hrs
			Experienced = 29.8 hrs

Tasks	3 simulated work tasks (fact, decision and background)	4 each of Medical, Buying, Travel and Project	3 simulated work tasks (fact, decision and background)
Experimental design	Graeco-Latin square	Latin square	Latin square
Tasks per subject	3	4	3

Time per task	10 minutes	10 minutes	10 minutes
Data Collection	Five questionnaires (One demographic, three system and one final) Background logging	Five questionnaires (One demographic and four system) Background logging	Five questionnaires (One demographic, three system and one final) Background logging Semi-structured interviews

In *TRSPresentation* I encourage subjects to employ other ways of examining search results, and use the sentence list as a replacement for the document list. In *TRSFeedback*, Top-Ranking Sentences were used to communicate system decisions in a comparison between implicit and explicit relevance feedback. *TRSDocument* uses the sentences to facilitate interaction with the top-ranked documents. The experimental system in *TRSDocument* still promotes the viewing of documents, but uses both documents and Top-Ranking Sentences. The content still drives the interaction with documents via the query-relevant sentences they contain.

The three studies are related and illustrate the initial stages of the development of my techniques. Top-Ranking sentences are first introduced as a replacement for document lists; I then study the substitutability of implicit and explicit feedback using these sentences. I finish by using both documents and sentences in a more intricate form of implicit feedback, based on the proof of substitutability that *TRSFeedback* provided me with. Figure 4.4 shows the relationship between the three user studies.



1. TRS as indicative of result set content
2. Shift focus from surrogate to content
3. TRS as **alternative** to document lists

1. TRS to reflect the use of relevance feedback: reordered in response to interaction

1. TRS to encourage document interaction
2. TRS as **complement** to document lists
3. TRS to reflect the use of relevance feedback: reordered in response to interaction

Figure 4.4. The relationship between the three user studies.

Top-Ranking Sentences drive searcher interaction. The same underlying motivation for their use applies in all three studies; ranking the content of the retrieved document set, rather than the documents themselves, helps subjects. In the next section qualitative results from the studies are presented and the implications of them discussed.

4.4 Findings and Discussion

In this section I present and discuss the qualitative findings of the user studies. The quantitative results, and more system details, have already been presented in White *et al.* (2003a) (*TRSPresentation*), White *et al.* (2002b) (*TRSFeedback*) and White *et al.* (2002a) (*TRSDocument*). Since the studies were conducted with different subjects, on different systems, at different times, direct comparisons across studies is difficult. Therefore I focus mainly on subject opinions of the search process, the Top-Ranking Sentences and the implicit feedback used to reorder the sentences. The findings discussed motivate the systems developed in the remainder of this thesis.

4.4.1 Search Process

Kuhlthau (1991) introduced a six-stage model of the Information Search Process (ISP), where searchers seek meaning from information to enhance their knowledge of their current problem or search topic. In this section, where appropriate, I discuss the findings of the user studies in relation to this model.

The experimental systems described in this chapter present a large amount of information at the search interface. There were concerns that this information would hinder subjects and lead to cognitive overload. In cognitive overload situations, a searcher's finite cognitive resources are stretched ever thinner by increased demands placed on them to process information. When faced with a plentiful supply of information, searchers typically have to make a series of decisions: Is this title relevant? Are these terms in the correct context? What comes after the ellipses? Where are these snippets in the document? Is the surrogate relevant? Shall I click this title? Every decision has an associated cost: time, effort and stress (Kirsh, 2000). The Top-Ranking Sentences restrict the decisions searchers make to those about the *relevance* of the information: Is this sentence relevant? Shall I click the sentence?

Subjects in all studies were asked to comment on the search process they performed on each of the systems, in particular they were asked how *stressful/relaxing* the search process had been. Cognitive overload scenarios can create *information anxiety* (Wurman, 1989) where the searcher becomes overwhelmed by information and trapped between their current state of knowledge and the amount of knowledge they require to solve the problem that initiated their seeking. Kuhlthau (1991) suggests that anxiety is an intrinsic part of the search process and will not totally disappear until the subject has successfully completed their task. However, it is possible to minimise this anxiety by providing levels of support that help subjects reach their goal. In the three studies, the presentation of more content at the results interface did not lead to high levels of stress reported by subjects during their search; generally subjects found the experimental systems intuitive. This is a worthwhile finding, as the benefits of Top-Ranking Sentences could be nullified if subjects felt stressed using the systems.

Kuhlthau's model of the ISP is divided into six stages that describe the search from beginning to end: initiation, selection, exploration, formulation, collection and presentation. Each stage has common affective, cognitive and physical activities and requires different levels of support from a search system. The systems described in this chapter support three of the six stages: *exploration*, *formulation* and *collection*. The other stages are typically carried out before the search system is used (understanding their information need and selecting search topics) or after the conclusion of the search (reporting the findings).

During the *exploration* stage subjects try to find information that will increase their understanding of what information is needed to complete their search. Kuhlthau (1991) suggests that during the exploration stage, strategies "…which open opportunities for forming new constructs such as listing facts which seem particularly pertinent…may be helpful during this time". The Top-Ranking Sentences are a list of query-relevant document representations that may help subjects better understand their information need and begin conceptualising these needs to form search statements.

The systems presented in this chapter provide limited support for the *formulation* stage of the ISP. This assumes that there is a point of 'focus' (Kelly, 1963; Belkin, 1980; Kuhlthau, 1991) where uncertainty drops and searchers can better identify the topic of their search. During this stage searchers formulate a focus during which they better understand their information need and the information they are searching for. The formulation stage is personalised and search systems that fully support it help searchers construct query statements. In the systems described in this chapter it is the system's internal representation of the information need that changes when presented with relevance information. This is hidden from the searcher, who only sees the effect of the revised formulation i.e., the reordered list of Top-Ranking Sentences. The systems support the improvement of search queries but since there is no direct dialogue with the searcher about these new queries their support for the formulation stage of the ISP is limited.

The experimental systems may also be useful during the *collection* stage of the ISP. The presentation of Top-Ranking Sentences gives searchers an opportunity to examine search results more closely and gather pertinent information from a variety of information sources. The search statements created as 'focus' was obtained are improved and enhanced (internally) and used to reorder the top-ranking sentence lists during the search. In the next section I discuss subject perceptions of the Top-Ranking Sentences.

4.4.2 Top-Ranking Sentences

The Top-Ranking Sentences were generally well received by experimental subjects. Although, from the user studies it did emerge that the training task and orientation sessions were important as subjects initially expressed concerns about the unfamiliarity of the interface. In this section I discuss subject perceptions of the TRS under three main section headings: task, popularity and usability.

4.4.2.1 Task

There were variations in the performance of top-ranking sentence based interfaces for different types of search task in the *TRSPresentation* and *TRSDocument* studies. Subjects felt that *background* and *decision* tasks required information from a number of sources to get a general overview of a topic or to make reasonable search decisions. The Top-Ranking Sentences were effective at facilitating access to such information. However, for the *fact* searches the Top-Ranking Sentences were not perceived as being as useful. That is, when searchers were fully aware of what they were looking for, they felt that they did not require additional interface support, and that they would be best able to find useful information with the commercial search engine they used most frequently. This does not imply that the Top-Ranking Sentences were useless; they were simply not required for the completion of this type of search task.

4.4.2.2 Popularity

Any problems experienced by subjects were mainly related to their unfamiliarity with top-ranking sentence-based interfaces. To interact well with the systems presented in these studies subjects had to change the way they searched for useful information. The approach encouraged more examination of search results and a reduction in the number of query reformulations; a shift from the well-established search paradigm currently promoted by commercial Web search engines. The negative findings above do not express a dislike for Top-Ranking Sentences, but for *any* change in the way results are presented. This may also suggest that if subjects are confident about being able to find information before starting to search they would rather use a familiar system (i.e., one where they do not have to think much about the interaction or the interface itself).

The value of titles, sentence fragments and URLs used by traditional Web search engines were tested in *TRSPresentation*. Subjects use these surrogates to make decisions about which documents to download and view. The user studies demonstrated that subjects rarely use interface features such as the 'next' button (all studies) or the URL of the document (*TRSPresentation* [13]). In the top-ranking sentence systems the URL and the 'next' button, although present, were not regarded as being as important.

Across all studies, the sentences and associated interface features were liked by subjects. In *TRSPresentation* I shifted the focus from document surrogates to the actual content of the document. In doing this, I found that the document titles were less useful as subject attention was drawn to the information resident inside documents. The experimental system used in *TRSPresentation* increased awareness of returned document set content, allowing subjects to make better decisions on the relevance of both the retrieved set of documents and documents individually.

4.4.2.3 Usability

In the experimental systems that presented results as a ranked list of documents subjects would rather reformulate and resubmit their queries than deeply peruse the documents returned to them. In doing so they may discard potentially relevant documents without giving them due consideration. The document list returned is only an algorithmic match to the searcher's query, something that typically contains only one or two query terms (Jansen *et al.*, 2000). Unless the information need is very specific (i.e., someone's name, such as in the *fact* search) the system may struggle to provide a ranking that is a match for the searcher's information need. This problem is amplified if the system only ranks whole documents as small highly relevant sections may reside in documents with a low overall ranking.

The Top-Ranking Sentences encourage more interaction with the retrieved document set, lowered the number of queries submitted and improved task success. Table 4.4 shows the percentage differences with the experimental baselines ($S_{Baseline}$, $S_{TRSAbstract}$ and $S_{Summarisation}$) used in the *TRSPresentation* and *TRSDocument* studies. If more than one top-ranking sentence system is used in the study or there is more than one non-TRS baseline then results are averaged across systems. All differences reported in the table were significant at $p < .05$.

[13] This was the only study where I measured the usefulness of the URL.

Table 4.4
Percentage difference between TRS systems and experimental (ranked document) baselines.

	Experimental factor				
	Page views			Task completion	
Study	Overall	Outside first 10	Queries	Time	Number of Tasks
TRSDocument	+ 43.59	+ 76.46	– 38.80	– 8.50	+ 16.67
TRSPresentation	+ 65.41	+ 115.44	– 61.20	– 8.68	+ 18.32

As can be seen from Table 4.4, the Top-Ranking Sentences encourage more page views outside the top 10 documents, more page views in general and a reduced number of query iterations. The increased number of page views coincided with a greater sense of task completion. The reduced number of queries suggests that subjects were interacting in a way symptomatic of increased perusal with the retrieved document set. The shorter task completion times and increased number of tasks completed suggests that subjects were using their time more efficiently. In the next section I discuss the results obtained on the implementation of implicit feedback in the experimental systems.

4.4.3 Implicit Feedback

The traditional view of information seeking assumes a searcher's need is static and represented by a single query submitted at the start of the search session. However, as is suggested by Harter (1992) among others, the need is in fact dynamic and changes to reflect the information viewed during a search. As they view this content their knowledge changes and so does their problematic situation. It is therefore preferable to express this modified problem with a revised query. The experimental systems in *TRSFeedback* and *TRSDocument* do this, selecting the most useful query expansion terms during a search.

In the systems developed in these studies, the sentences are reordered using implicit relevance information gathered unobtrusively from searcher interaction. Experimental subjects found this a useful feature that helped them find relevant information. They suggested that it was most useful when they felt the initial query had retrieved a large amount of potentially relevant information and they wanted to focus their attention on only the most relevant parts. These are more push oriented than the static Top-Ranking Sentences system tested in *TRSPresentation*. The systems are adaptive, work to better represent information needs and consider changes in these needs, restructuring the content presented at the results interface.

In *TRSFeedback* and *TRSDocument* I assumed that the viewing of a document's summary was an indication of an interest in the relevance of the summary's contents. There are several grounds on which this can be criticised; searchers will view non-relevant summaries, the title rather than the summary was what the user expressed an interest in, and the searcher may look at all retrieved documents before making real relevance decisions. Nevertheless I felt that this assumption was fair enough to allow an initial investigation into the use of implicit feedback. In *TRSDocument* I introduced a timing mechanism to eliminate the problems caused by the accidental 'mouseover' of document titles and the unwanted removal of sentences from the Top-Ranking Sentences list that follows. The results of *TRSDocument* are testament to the success of a very limited version of an implicit feedback technique. More complex and effective techniques based on these findings are described in later chapters of this thesis.

Despite their positive comments, subjects had two reservations about how system decisions based on implicit feedback were communicated. Firstly, since the reordering occurred at the same time as a summary appeared or updated they did not always notice the effect of the reordering. The presentation of the updating therefore needs improving in future systems. Secondly, the Top-Ranking Sentences only contained sentences from Web pages for which the subject had not already viewed a summary. If the subject viewed the summary for a page, then all sentences from that page were removed from the list of Top-Ranking Sentences. This choice was made to increase the degree to which the list of Top-Ranking Sentences would update. However, many subjects stated that they would prefer less updating and no removal of sentences. In White (2004) I proposed the use of *ScrollTiles* to communicate the effects of the sentence reordering using a familiar interface component, the scrollbar. The approach represented sentences as tiles on the scrollbar and re-coloured the tiles to represent changes in the ordering. A pilot study was conducted that involved nine experimental subjects and compared systems that re-coloured a representation of the sentences imposed on the scrollbar with one that reordered the actual sentences. The ScrollTiles were shown to be more effective for conveying reordering decisions than the sentence updating. However, they are not used in any further interfaces described in this thesis as I tried to limit the number of new interface components to only those necessary to test experimental hypotheses.

The results of the three studies show that it is possible to get searchers to interact with more than a few search results. The approach moves away from simply presenting titles to presenting alternative access methods for assessing and targeting potentially relevant information. The findings were useful in the development of search interfaces described later in this thesis.

4.5 Summary

Ranking documents is potentially a heavy-handed, cumbersome means of result presentation. Documents may not be entirely relevant and document surrogates may not be strictly indicative; it is the information in the documents that searchers seek. The content-driven approach extracts, ranks and presents the content of the returned set, blurring inter-document boundaries and encouraging information seeking based on the potentially relevant document content.

In this chapter I have discussed the results of three studies to test the effectiveness of content-driven information seeking. The implicit feedback frameworks proposed in this thesis rely on searcher interaction with the retrieved information as evidence of what information is relevant. The studies presented in this chapter show that the interfaces developed are liked by subjects and can lead to more effective information seeking. This was a promising finding for the development of search systems developed later in this thesis. The studies have also highlighted problems in the use of these interfaces that are addressed in later systems. In Chapter Five I present an overview of a search interface that uses titles, summaries and Top-Ranking Sentences and other document representations to facilitate access to potentially relevant information.

Chapter 5

Representations and the Search Interface

5.1 Introduction

In this chapter I describe the document representations presented at the search interface and used by the implicit feedback frameworks described in the thesis. These representations are typically sentence-based and created by the search system at retrieval time. As they are small, interaction with document representations is potentially more focused than with the full-text of documents and since they are numerous, can generate an increased quantity of evidence for the implicit feedback frameworks. In Chapter Four document representations were used to encourage searchers to interact more with the results of their search. Through presenting multiple representations of the same document it is possible for searchers to directly indicate which document components (e.g., sentences, summaries, and titles) are relevant. Traditional RF techniques rely on searcher feedback about the relevance of whole documents; this can be unreliable as documents can contain irrelevant parts. The principle of *polyrepresentation* (Ingwersen, 1994) suggests that IR systems should provide and use different cognitive structures during acts of communication to reduce the uncertainty associated with interactive IR. The techniques I describe implement one aspect of a polyrepresentative approach; the use of multiple document representations.

The chapter also presents the generic design of a search interface that combines document representations in an interactive context. The document representations and interface presentation techniques are described in the remainder of this chapter.

5.2 Document representations

IR systems were originally designed for the retrieval of documents from homogeneous corpora, such as newspaper collections or library index cards. Document surrogates, such as titles and abstracts, were usually created by experts, such as librarians or professional cataloguers. The growth in size, dynamism and heterogeneity of the collections being searched led to the development of automated representation techniques and a reduction in the quality of the surrogates created. However, work by Landow (1987) and Furnas (1997) has shown the importance of the information that searchers use when deciding which documents to download and view. If the quality of document representations has decreased, then one possible solution is to increase the quantity of information available to view. That is, provide searchers with more information to make search decisions.

In my approach the most relevant documents in the retrieved set are represented by a variety of document representations. The principle of polyrepresentation (Ingwersen, 1994) suggests that different cognitive structures should be offered to searchers and used by them during their interaction with an 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 different *cognitive origins*. The author's text, including titles and the full-text are representations of cognitive structures intended to be communicated. However, these portions of text have different *functional origins*. That is, they have the same cognitive origin but were created in a different way or for a different purpose.

In Chapter Four experimental search interfaces were presented that used different representations of the top-ranked documents. In those studies Top-Ranking Sentences, titles and document summaries were used to represent their source documents and facilitate effective information access. In this chapter, three further representations are used: summary sentences, summary sentences in document context and the full-text of the document. These representations describe the document in different ways. The full-text is only the textual content of the document; all other document features, such as images and document structure, are ignored since they cannot be used by the sentence selection methods described in Chapter Three.

The sentence-based representations (i.e., Top-Ranking Sentences, 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). These representations are created using algorithms devised by the system designers and are selected based on queries submitted by a searcher, both cognitive agents. Offering searchers different representations of the same document at the search interface is one aspect of polyrepresentation. However, the basis of polyrepresentation is the use of the overlap between these representations to reduce uncertainty. The theory has been implemented across networks of citations (Larsen and Ingwersen, 2002), where those who cite documents have unique cognitive structures. The interface described in this chapter use many document representations to implement one aspect of a polyrepresentative approach that aim to reduce the uncertainty associated with gathering implicit feedback. In this section I introduce each of the representations and explain their role in the search interface.

5.2.1 Top-Ranking Sentences

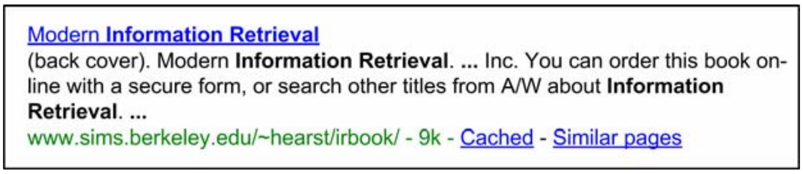
Top-Ranking Sentences were introduced in Chapters Three and Four as a means of facilitating access with retrieved information. The results of the user studies in Chapter Four demonstrated the usefulness of presenting sentences in a list, ranked independently of their source documents. The interfaces described in this chapter use these sentences in the same way. Ingwersen (1994) 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 (Salton *et al.*, 1993; Callan, 1994). In the search interfaces I create, the Top-Ranking Sentences provide a starting point from which searchers can access potentially useful information. The sentences may contain the information necessary to satisfy their information need, or may provide a means through which searchers can access relevant documents.

5.2.2 Document Title

This is the title of the document, as assigned by the author. Titles are typically short and include terms that express the main themes of a document. On the Web, the corpus for the user studies described in this thesis, authors assign document titles and the extent to which they are indicative of current document content can vary.

5.2.3 Document Summary

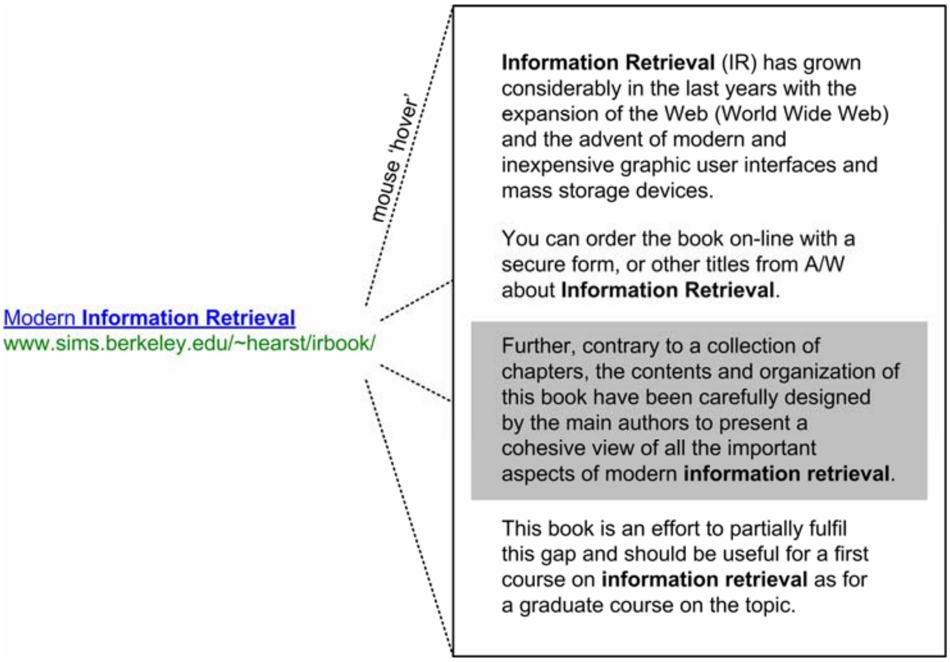
A document summary contains the four Top-Ranking Sentences for that document. The summary is based on the query submitted by the searcher and is created in real-time, when a query is submitted, using the best Top-Ranking Sentences selected by the approach given in Chapter Three. Figure 5.1 shows an example summary produced by the Google Web search engine. This summary is typically composed of a series of sentence fragments that could contain the query terms, separated by ellipses.



Modern **Information Retrieval**
(back cover). Modern **Information Retrieval**. ... Inc. You can order this book on-line with a secure form, or search other titles from A/W about **Information Retrieval**. ...
www.sims.berkeley.edu/~hearst/irbook/ - 9k - Cached - Similar pages

Figure 5.1. Document abstract from Web search engine for query 'information retrieval'.

Figure 5.2 shows the summary generated by combining the same document's four best Top-Ranking Sentences. The summary window on the right of the figure appears immediately or after a short time delay when the searcher hovers over the document title.



mouse 'hover'
Modern Information Retrieval
www.sims.berkeley.edu/~hearst/irbook/

Information Retrieval (IR) has grown considerably in the last years with the expansion of the Web (World Wide Web) and the advent of modern and inexpensive graphic user interfaces and mass storage devices.

You can order the book on-line with a secure form, or other titles from A/W about **Information Retrieval**.

Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors to present a cohesive view of all the important aspects of modern **information retrieval**.

This book is an effort to partially fulfil this gap and should be useful for a first course on **information retrieval** as for a graduate course on the topic.

Figure 5.2. Document summary from the best four Top-Ranking Sentences for query 'information retrieval'.

The difference in the content and quality of the summaries between the two summary generation approaches is significant. The summaries created by combining the best Top-Ranking Sentences are semantically richer and may allow more accurate relevance assessments than standard search engine summaries (White *et al.*, 2003b).

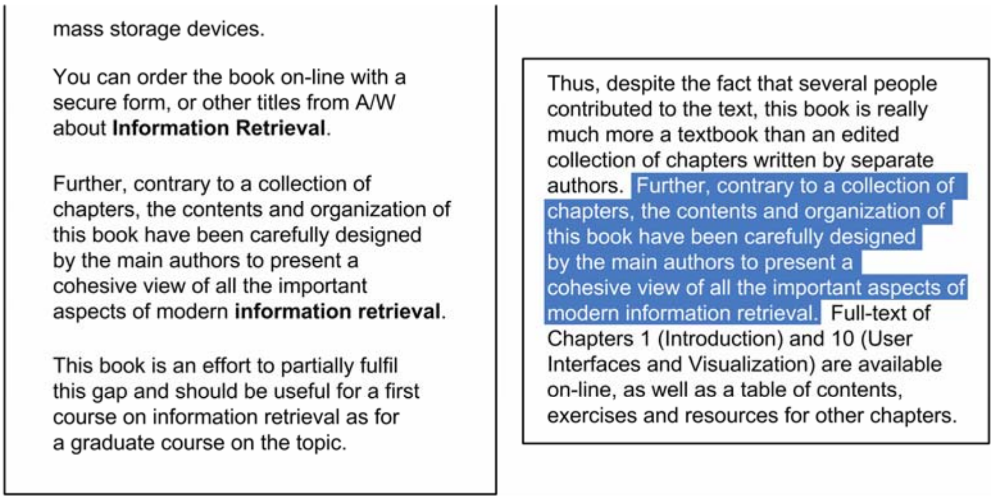
5.2.4 Summary Sentence

Each sentence in the summary of the document is considered a representation of the source document. Allowing relevance assessments at the sentence level allows for more precise assessments of what information meets searchers needs. In Figure 5.2 the third summary sentence is highlighted.

5.2.5 Sentence in Context

A summary sentence in the context in which it occurs in the document (i.e., preceding and following sentence from the source document) is also available for searchers to view. This 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, if there are the two sentences: “Alexander Graham Bell invented the telephone. He emigrated to Canada when he was just 23”. 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 be problematic if the latter sentence is shown without the first. Presenting the latter sentence in the original document context can contribute to the resolution of such problems.

In Figure 5.3 the highlighted sentence in Figure 5.2 is shown in the context in which it occurs in the source document. The sentence in context appears directly next to the sentence in summary to make the association between the two representations more clear. In the ‘Sentence in Context’ window on the right of Figure 5.3, the summary sentence is highlighted and the preceding and following sentences are also shown to the searcher.



mass storage devices.

You can order the book on-line with a secure form, or other titles from A/W about **Information Retrieval**.

Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors to present a cohesive view of all the important aspects of modern **information retrieval**.

This book is an effort to partially fulfil this gap and should be useful for a first course on information retrieval as for a graduate course on the topic.

Thus, despite the fact that several people contributed to the text, this book is really much more a textbook than an edited collection of chapters written by separate authors. Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors to present a cohesive view of all the important aspects of modern information retrieval. Full-text of Chapters 1 (Introduction) and 10 (User Interfaces and Visualization) are available on-line, as well as a table of contents, exercises and resources for other chapters.

Figure 5.3. Summary sentence in document context.

The sentences in context are created immediately after the retrieved documents have been summarised (i.e., after query submission and before result presentation). Figure 5.4 shows the process involved to create the sentence in context for each sentence in the document summary. First the Top-Ranking Sentences are selected from the source document, and the sentences that comprise the summary are passed to the context generation component. Each sentence has a unique identifier, based on its position in the document. The context generation component then locates the sentence that immediately precedes and immediately follows the summary sentence. For example, in Figure 5.4 sentence s_3 is a summary sentence and sentences s_2 and sentence s_4 form the context for s_3.

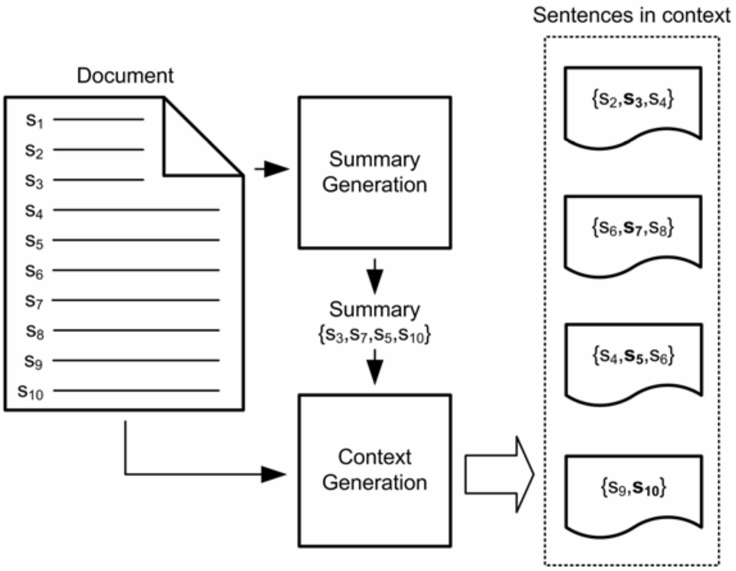


Figure 5.4. Creation of sentence in document context.

If a sentence is the last sentence in a document (as with s_{10} in Figure 5.4) only the sentence before is used to compose the sentence in context. Since s_{10} is the last sentence, the context will only comprise s_{10} and the prior sentence s_9. The same is true for the first sentence, except that the only the sentence directly following it is used to comprise the context.

5.2.6 Document (Full-text)

The full-text is the document, as created by the author. The full-text of the document is the source of the sentences used to create the document representations. Monitoring searcher interaction with documents is problematic as it can be difficult to determine exactly what part of the document, if any, searchers regard as relevant. Using all terms from documents searchers view may adversely affect the retrieval performance of the term selection parts of the model (Salton *et al.*, 1993), especially if the document is actually irrelevant. Therefore, the document full-text is not used directly in any of the implicit feedback frameworks described in this thesis. However, the set of terms extracted from the set of most relevant documents forms the vocabulary or term space used by the implicit feedback frameworks described in later chapters.

5.2.7 Overview of Representations

There is redundancy in the representations that searchers interact with. A single top-ranking sentence may appear in five of the six document representations: the Top-Ranking Sentences list, the document summary, a summary sentence, a sentence in context and the source document. Searcher interaction with the same sentence in a number of representations provides more evidence for the relevance of representations.

Different types of representation vary in length, and can hence be regarded as being more or less *indicative* of the content of the document (Barry, 1998). 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. The *length hypothesis* (Marcus *et al.*, 1978) suggests that the quality of a representation is directly proportional to its length. The validity of this hypothesis had been supported by previous work (Weis and Katter, 1967; Hagerty, 1967). However, the hypothesis has been criticised for failing to consider the quality or nature of a representation (Janes, 1991). For example, a document title is typically short but is assigned by the author, and may capture the key concepts in a document. The heuristic-based implicit feedback framework described in Chapter Six uses the length hypothesis to assign an indicativity weight to the representations.

An alternative approach is to assume that representations that are more indicative of their source documents contribute more to the refinement of query statements. Janes (1991) views the length hypothesis as superficial and perhaps more suited for heuristic-based approaches. The probabilistic implicit feedback framework presented in Chapter Seven does not use representation length as a measure of representation quality. Instead, it gives more weight to representations with higher quality content. To do this, it constructs an *indicativity index* (White *et al.*, 2004b) measured based on the terms that co-occur between the representation and the document. Representations that are highly indicative of the source document are regarded as high quality. Some representations of each document are fixed in content, i.e., the title and full-text of the document, whereas other representations, such as the summary, are dependent on the query and hence variable in content. The document title and the full-text are created by the author and are not query dependent.

In the next section I describe the search interface that combines the document representations for the presentation of search results.

5.3 Search Interface

The search interface presents a variety of document representations to the searcher. These *content-rich* search interfaces present more information from retrieved documents than standard search engine interfaces. Through their interaction searchers can control which representations are shown on the interface at any one time. A schematic of the interface is shown in Figure 5.5. The 'Summary', 'Sentence in Context' and 'Document full-text' all become the active window – displayed in front of the other information – when the searcher requests them. The default display is the list of Top-Ranking Sentences and the list of document titles. The list of Top-Ranking Sentences can contain around 60 sentences from the most relevant Web documents.

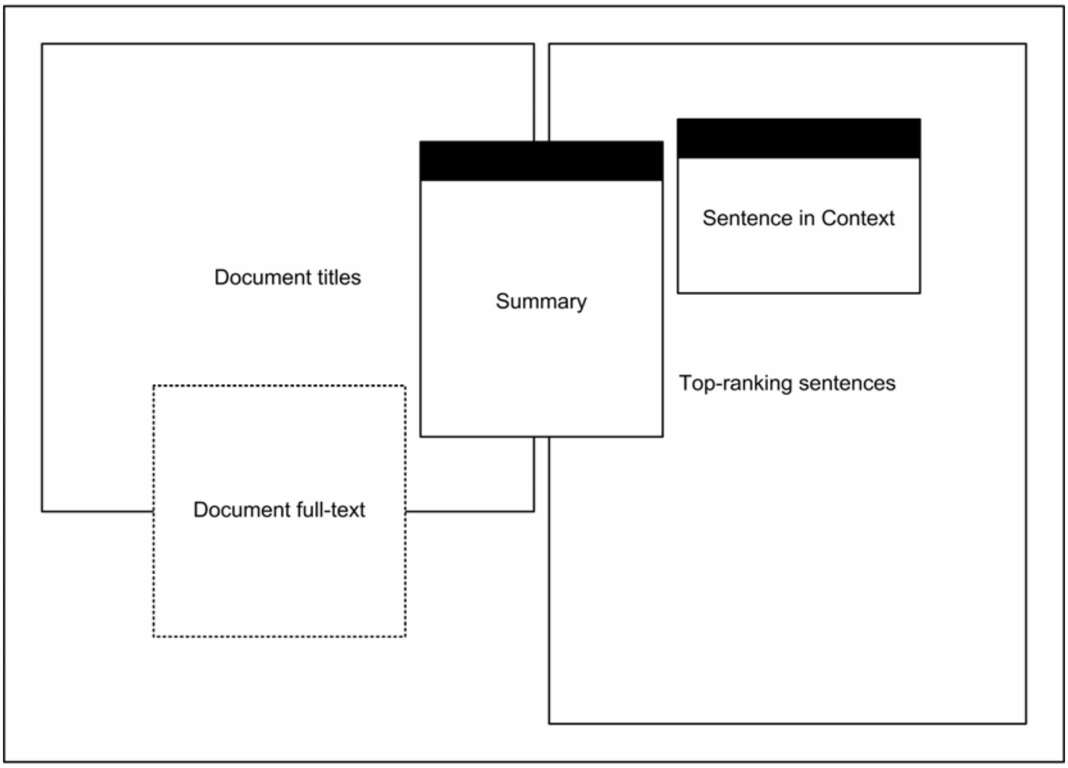
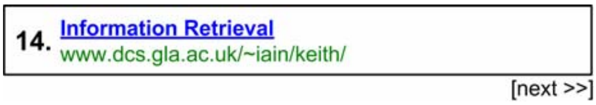


Figure 5.5. Schematic of the search interface.

This style of interface was chosen since it allows the search system to closely monitor what document representations searchers may be viewing at any given time. This allows implicit feedback frameworks that use interaction with these interfaces to make potentially more accurate inferences about searcher interests. Searchers can view the title of a top-ranking sentence's source document simply by interacting with the sentence. Should the title fall outside the first 10 documents then a small window below the list of document titles updates to show the title (as a clickable hyperlink) and in some systems the URL. An example of this window is given in Figure 5.6.

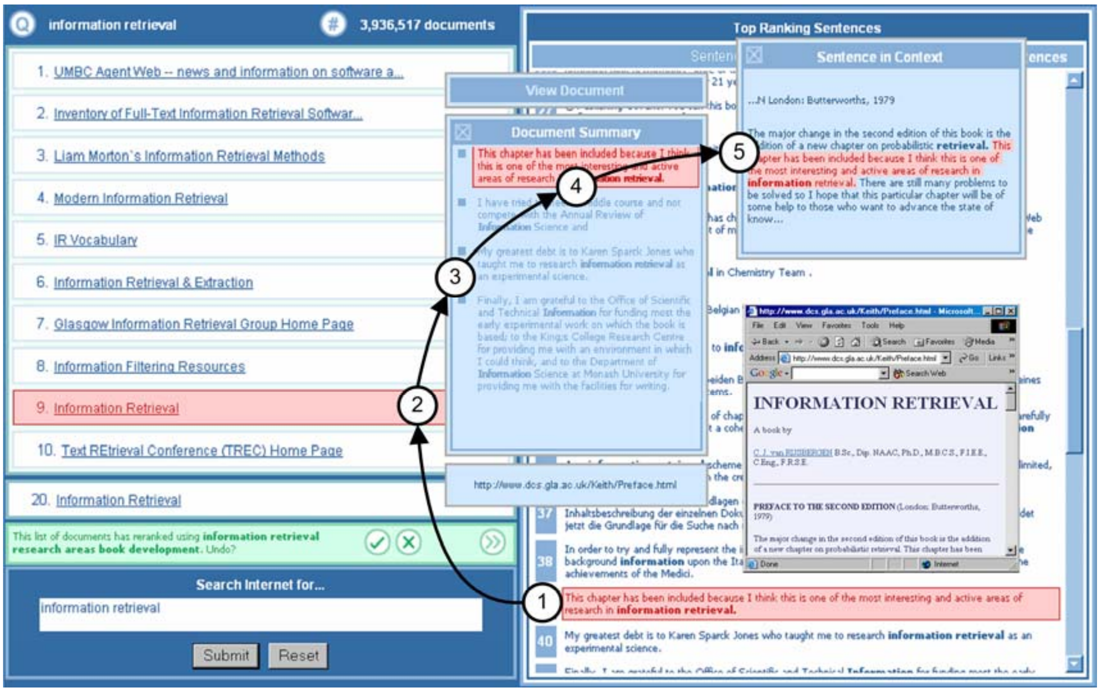


14. Information Retrieval
www.dcs.gla.ac.uk/~iain/keith/

[next >>]

Figure 5.6. Document title pop-up for documents outside the top ten retrieved.

Searchers can interact with the hyperlink in this window in the same way as with any title in the first 10 retrieved documents. That is, they can click the text to visit the document or hover over the title to see a summary of the document. Figure 5.7 shows an experimental interface used in Pilot Test 1, described in Chapter Nine, which implements these concepts.



information retrieval 3,936,517 documents

1. UMBC AgentWeb -- news and information on software a...
2. Inventory of Full-Text Information Retrieval Softwar...
3. Liam Morton's Information Retrieval Methods
4. Modern Information Retrieval
5. IR Vocabulary
6. Information Retrieval & Extraction
7. Glasgow Information Retrieval Group Home Page
8. Information Filtering Resources
9. Information Retrieval
10. Text REtrieval Conference (TREC) Home Page

20. Information Retrieval

This list of documents has reranked using **information retrieval research areas book development**. Undo?

Search Internet for...

information retrieval

Submit Reset

Document Summary

- This chapter has been included because I think this is one of the most interesting and active areas of research in information retrieval.
- My greatest debt is to Karen Sparck Jones who taught me to research **information retrieval** as an experimental science.
- Finally, I am grateful to the Office of Scientific and Technical **Information** for funding most the early experimental work on which the book is based; to the Kings College Research Centre for providing me with an environment in which I could think, and to the Department of **Information** Science at Monash University for providing me with the facilities for writing.

http://www.dcs.gla.ac.uk/Keith/Preface.html

Sentence in Context

...N London: Butterworths, 1979

The major change in the second edition of this book is the addition of a new chapter on probabilistic **retrieval.** This chapter has been included because I think this is one of the most interesting and active areas of research in **information** retrieval. There are still many problems to be solved so I hope that this particular chapter will be of some help to those who want to advance the state of know...

http://www.dcs.gla.ac.uk/Keith/Preface.html - Microsoft...

INFORMATION RETRIEVAL

A book by

C.J. van RIJSBERGEN B.Sc., Dip. NAAC, Ph.D., M.B.C.S., F.I.E.E., C.Eng., F.R.S.E.

PREFACE TO THE SECOND EDITION (London: Butterworths, 1979)

The major change in the second edition of this book is the addition of a new chapter on probabilistic retrieval. This chapter has been

Top Ranking Sentences

37 Inhaltsbeschreibung der einzelnen Doku... jetzt die Grundlage für die Suche nach...

38 In order to try and fully represent the i... background **information** upon the Ita... achievements of the Medici.

This chapter has been included because I think this is one of the most interesting and active areas of research in **information retrieval.**

40 My greatest debt is to Karen Sparck Jones who taught me to research **information retrieval** as an experimental science.

Figure 5.7. Experimental search interface in Pilot Test 1 (Chapter Nine).

The effectiveness of top-ranking sentence-based interfaces to statically structure information spaces has already been demonstrated (Tombros *et al.*, 2003a; 2003b). In these studies Top-Ranking Sentences were clustered to create personalised search spaces that made interaction more effective. The implicit feedback frameworks described in this thesis modify the query and estimate changes in the information needs of searchers. Adaptive views of the information space can support the developing nature of information needs (Campbell, 1999). The frameworks restructure or recreate the search results at each query iteration to bring potentially relevant results to the attention of the searcher. The mechanisms behind the interface proposed in this section use searcher interaction to formulate a query that represents their information need and dynamically restructure or recreate the search results based on the predicted extent of any changes in this need.

Interacting in a certain way with each representation suggests another representation for that document. For each document it is possible to follow a path between its representations. These are called *relevance paths* since the further a searcher travels along a path the more evidence there is on the relevance of the path's resident information. Searchers are guided along the relevance path by their interaction and the search system. In the next section these paths are described.

5.4 Relevance Paths

There are many applications of paths in IR (Pirolli and Card, 1995; Campbell and Van Rijsbergen, 1996; Chalmers *et al.*, 1998). The Ostensive Model (Campbell and Van Rijsbergen, 1996) uses paths between documents or document representations to build a context for the search and choose appropriate terms to form a new query. Information foraging theory (Pirolli and Card, 1995) assumes users are driven by the to click hyperlinks based on proximal cues given by their surrounding text. The path model (Chalmers *et al.*, 1998) uses each individuals' ongoing history of ratings or choices to choose similar pages.

These applications all consider paths between documents e.g., clicking a hyperlink resident in one document to get to another document. However, the relevance paths I propose form *between document representations*. The paths provide searchers with progressively more information from the best documents to help them choose new query words and select what new information to view. The further along a path a searcher travels (i.e., the more representations in a path they view) the more relevant the information in the path is assumed to be. The order in which certain types of representation are available in a relevance path is dictated by the interface. Searchers are guided along the path by their interaction with the search interface. If they interact with the Top-Ranking Sentences the system highlights the title of the source document. If they hover over a document title for a short time the summary of that document appears in a small, moveable window in front of the other information. Clicking arrows next to sentences in that summary shows the sentences in the context they occur in the source document.

The paths can vary in length from one to six representations long, and searchers can access the full-text of the document from any step in the path by clicking on the text of the document representation. Since searchers can take many routes between representations for each document, there may be many *potential* relevance paths. Relevance paths can start from Top-Ranking Sentences or document titles. Certain aspects of the path order are fixed e.g., the searcher must view a summary sentence before visiting that sentence in context. The full-text of the document is accessible from all representations. That is, a searcher can click on all representations and access the source document. There are 54 potential relevance paths for each document. In Figure 5.8 I show a possible relevance path route for a single document, at each step a representation is viewed (shown in darker font).

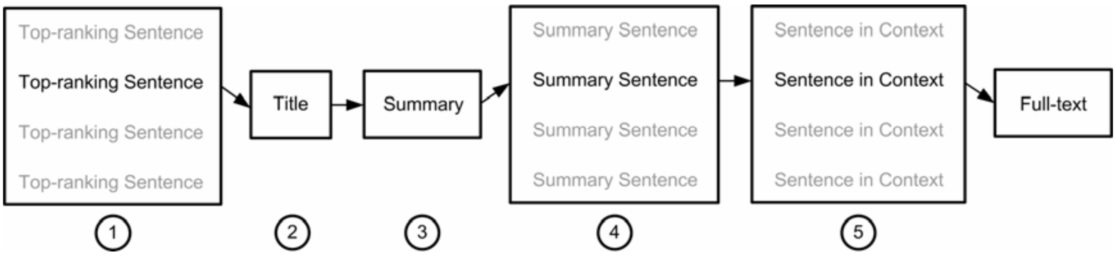


Figure 5.8. Possible relevance path route (numbers correspond to Figure 5.7).

In Figure 5.8 a top-ranking sentence is viewed, followed by the title of the document, the summary for that document, a sentence in that summary and in context, followed by the full-text of the document. There are six steps in this relevance path. In Figure 5.9 this relevance path is shown on the interface schematic. To follow this path a searcher would have to interact with each of the representations on the path. The full-text of the path's source document is eventually accessed in this instance from the sentence in context.

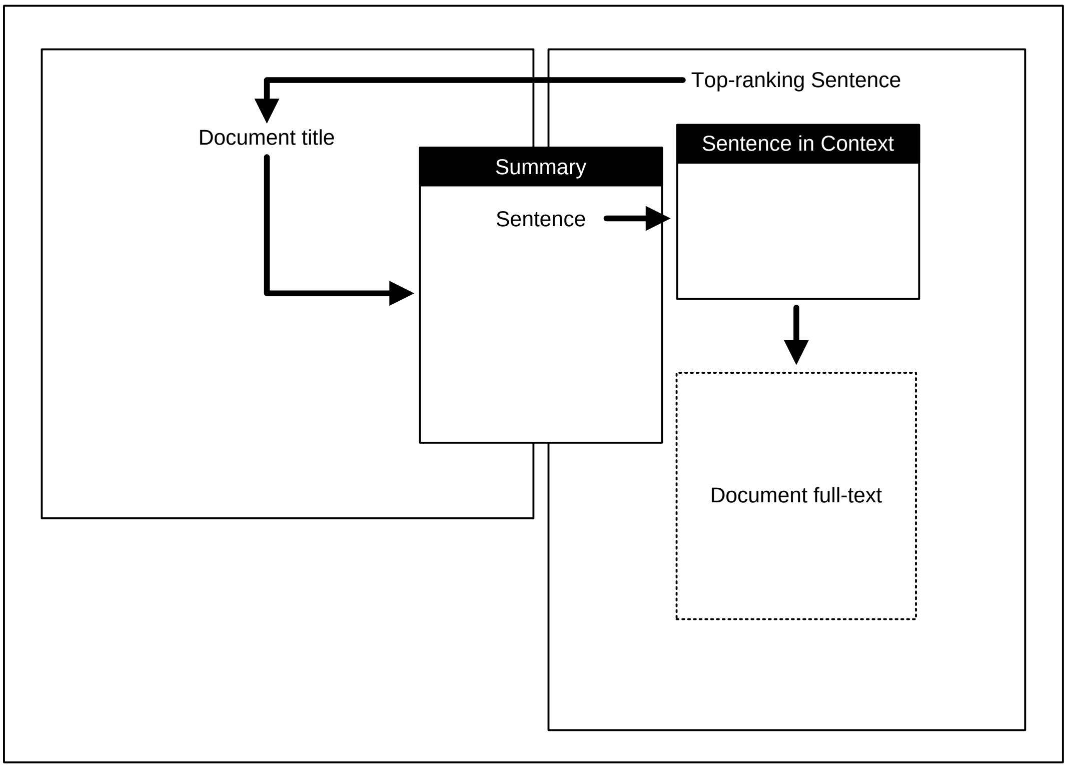


Figure 5.9. Possible relevance path on interface schematic.

As a searcher moves along the relevance path they move from assessing document representations in relation to other representations (i.e., Top-Ranking Sentences, titles) to a deeper examination of representations in their resident context (i.e., summaries, sentences in context). That is, as a searcher traverses a relevance path, their interaction with top-ranked documents becomes more focused. To the searcher, the path represents a desire to find out more information about a document or to find the information they require to satisfy their needs. To the implicit feedback framework operating behind the search interface, each relevance path is a source of evidence that allows it to build a body of relevance and make decisions on the searcher's behalf. Showing searchers progressively more information about a document to assist relevance assessments has already been used in related work (Zellweger *et al.*, 2000; Paek *et al.*, 2004).

5.5 Summary

In this chapter I have described the document representations presented to searchers at the interfaces described in this thesis. These representations allow searchers to view and assess the relevance of information at the results interface rather than visiting documents and locating the information inside them.

Document representations are linked at the interface by relevance paths that guide searcher interaction. The further along a relevance path a searcher travels, the more relevant the information in the path is assumed to be. These paths are included in the content-rich interfaces described in this chapter and aim to encourage searchers to interact with the retrieved information in a structured way, generating more evidence for the implicit feedback frameworks that use this as evidence of searcher interests. In forthcoming chapters the frameworks that utilise this interaction are presented.

Part III

Implicit Feedback Frameworks

In Part II I described content-driven techniques that use a variety of document representations to facilitate an increase in the quality and quantity of interaction with retrieval systems. Techniques for selecting query-relevant Top-Ranking Sentences were presented and the results of three studies that tested the effectiveness of the approach were described. The success of the implicit feedback techniques described in Chapter Four encouraged me to enhance these methods. In this part I present two implicit feedback frameworks that use interaction with the interface described in Chapter Five to infer information needs and select retrieval strategies. One of the frameworks is heuristic-based and the other is probabilistic. Part III concludes with a novel simulation-based comparative evaluation of the term selection models in the frameworks and other models. The simulation emulates searcher interaction with content-rich interfaces and serves a formative evaluation technique to establish the most effective implicit feedback model to be tested in later experiments.

Chapter 6

Heuristic-Based Framework

6.1 Introduction

In this chapter the first of two implicit feedback frameworks introduced in this thesis is described. Both frameworks use unobtrusive monitoring of interaction to proactively support searchers. The framework chooses terms to better represent information needs by monitoring searcher interaction with different representations of top-ranked documents. As suggested in Chapter Two, information needs are dynamic and can change as a searcher views information. The framework proposed gathers evidence on potential changes in these needs through changes in term lists used for query formulation by the search system. The framework uses the evidence it gathers to choose new retrieval strategies such as re-searching the document collection or restructuring already retrieved information. Large estimated changes in information need lead to more severe interface support. The framework described in this chapter is heuristic-based and uses term presence/absence in viewed representations to select terms for query modification.

6.2 Information Need Detection

In this section I describe the *Binary Voting Model*, a heuristic-based implicit feedback model I develop to implicitly select terms for query modification. The approach utilises searcher interaction with the document representations and relevance paths described in the previous chapter. The representations viewed by a searcher are used to select new query terms and in the Binary Voting Model each representation ‘votes’ for the terms it contains. When a term is present in a viewed representation it receives a ‘vote’, when it is not present it receives no vote. [14] All non-stopword, non-stemmed terms in the top-ranked documents are candidates in the voting process; these votes accumulate across all viewed representations. The assertion I make is that the winning terms are those with the most votes, and hence best describe the information viewed by the searcher. I assume that useful terms will be those contained in many of the representations that the searcher chooses to view. The rationale behind this assertion is that searchers will try to maximise the amount of relevant information they view during a search (Pirolli and Card, 1995). The non-stopword terms that appear in the representations they view (and in similar contexts to their original query terms) are those that are potentially important to the searcher and may be useful for query modification.

6.2.1 Indicativity

In the Binary Voting Model terms are assigned a weight of one or zero, depending on whether they occur in a representation, regardless of the type of representation. All items presented to the searcher at the content-rich search interface are representations of the top-ranked documents. Different types of representation vary in length, and can hence be regarded as being more or less *indicative* of the content of the document. The *length hypothesis* (Marcus *et al.*, 1978) suggests that the quality of a representation is directly proportional to its length. That is, longer representations are regarded as being of a higher quality than shorter representations, simply because they reveal more of the document content. For example, a top-ranking sentence is less indicative (and therefore of a lower quality) than a query-biased document summary (typically composed of four sentences) as it contains less information about the content of the document. The model *weights* the contribution of a representation's vote based on its indicative worth. For example, I consider the contribution that viewing a top ranking sentence makes to the system's understanding of which terms are relevant to be less than a summary.

The weights used in the framework are 0.1 for title, 0.2 for Top-Ranking Sentence (TRS), 0.3 for Summary, 0.2 for Summary Sentence and 0.2 for Sentence in Context. For example all terms in a viewed summary will receive a weight of 0.3; all terms in a viewed summary sentence will receive a weight 0.2, etc. These weights were defined for experimental purposes and were based on the *typical* length of a representation, not their potential semantic

[14] The decision to use binary (term presence/absence in a representation) rather than term frequency (*tf*) information was taken for reasons of simplicity and computational expense. Through empirical investigation I tested the effectiveness of other methods of term weighting such as *tf*, *tf.idf*, *tf* normalised by representation length, none of which performed better than binary voting, and in the case of *tf.idf*, performed worse. This could be because it also included the importance of a term across all document representations relevant or not (i.e., the *idf* weight), not just those viewed.

value. They ensure that the total score for a term is between zero and one (inclusive) and are used in the absence of a more formal methodology.

6.2.2 Term Weighting

The model is a simple approach to a potentially complex problem. The terms with most votes are those that are taken to best describe the information viewed by the searcher (i.e., those terms that are present most often across all viewed representations) and can therefore be used to approximate searcher interests. Of course, searchers may view irrelevant information as they search. In general however, their interaction decisions are guided by a desire to maximise the amount of relevant information they view.

Each document is represented by a vector of length n; where n is the total number of unique non-stopword, non-stemmed terms in the top-ranked Web documents.[15] In this chapter the list holding these terms is referred to as the *vocabulary*. All terms in the vocabulary are candidates in the voting process.

To weight terms a document × term matrix, shown in Figure 6.1, $(d+1)\times n$ is constructed, where d is the number of documents for which the searcher has travelled at least part of the relevance path. Each row in the matrix represents all n terms in the vocabulary [i.e., $(t_{k1},t_{k2},\ldots,t_{kn})$ where k is the row number], and each term has a weight. An additional row is included for the query.

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

Figure 6.1. Document × Term matrix.

Query terms are initially assigned a weight of one if they are included in the query and zero if not. Example 6.1 (used throughout this chapter) illustrates the operation of the Binary Voting Model.

[15] I do not use word stems since they may not be interpretable by searchers unfamiliar with stemming.

Example 6.1: Simple Updating

If one assumes that there are only 10 terms in the vocabulary and that the original query (Q_0) contains t_5 and t_9, the document × term matrix initially looks like:

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

Each row in the matrix is normalised to give each term a value in the range [0, 1] and make the values sum to one. This ensures that the query terms are not weighted too highly in the document × term matrix. This is important when the model is *replacing* query terms; a high query term weight would lessen the chances of other terms being chosen. The matrix now looks like:

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

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

This weight for each term is *added* to the appropriate term/document entry in the matrix. Weighting terms is therefore a *cumulative* process; the weights calculated for a term in one representation are added to the weights calculated for the preceding steps in the relevance path. Unlike standard RF algorithms which calculate one set of weights for query modification terms between documents, the Binary Voting Model calculates weights on a per document basis (i.e., within documents). There are different sets of weights for each document and these weights correspond to a row in the document × term matrix.

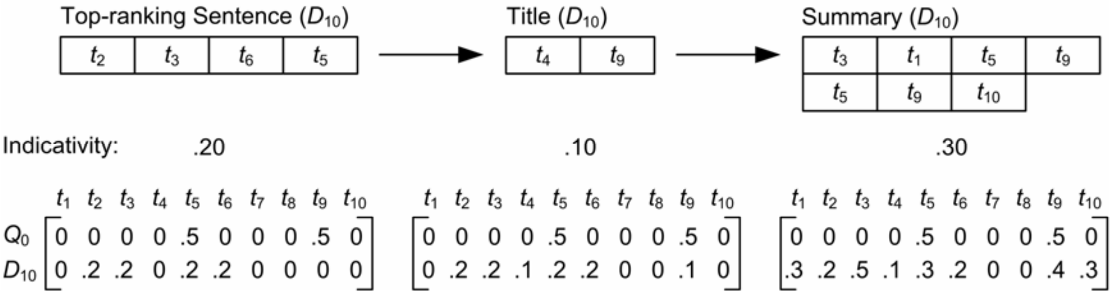
The total score for a term in a document is computed by:

$$w_{t,D} = \sum_{i=1}^{N} (w_{t,r}) \tag{6.1}$$

Where N is the number of steps taken in a relevance path, i is the current step number, D is the document, t is the term, r is the representation and $w_{t,r}$ is the weight of t for representation r.

Example 6.1: Simple Updating (continued)

When a searcher follows a relevance path, the model updates the weights in the document × matrix after each step. Figure 6.2 shows how the term weights are updated as a path from a top-ranking sentence, to title, to summary is traversed.



$$\begin{array}{c|cccccccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \\ \hline Q_0 & 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 \\ D_{10} & 0 & .2 & .2 & 0 & .2 & .2 & 0 & 0 & 0 & 0 \end{array} \quad \begin{array}{cccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \\ \hline 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 \\ 0 & .2 & .2 & .1 & .2 & .2 & 0 & 0 & .1 & 0 \end{array} \quad \begin{array}{cccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \\ \hline 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 \\ .3 & .2 & .5 & .1 & .3 & .2 & 0 & 0 & .4 & .3 \end{array}$$

Figure 6.2. Updating the Document × Term matrix.

The weights of all terms except t_7 and t_8 are directly updated. The terms whose weights update are seen as being more important than before to D_{10}. If the document × term matrix is as shown on the far right of Figure 6.2 and the searcher expresses an interest in the title of document D_5 – with a step weight of 0.1, containing terms t_3 and t_5 – the matrix changes to:

$$\begin{array}{c|cccccccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \\ \hline Q_0 & 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 \\ D_{10} & .3 & .2 & .5 & .1 & .3 & .2 & 0 & 0 & .4 & .3 \\ D_5 & 0 & 0 & .1 & 0 & .1 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Figure 6.3. Document × Term matrix after addition of new document, D_5.

If the searcher visits one representation of a document and then goes onto the next representation in the path of that document, *at any time – not necessarily immediately*, the model adds the term scores to the row in the matrix occupied by that document. The scoring is cumulative; if a document already has a row in the matrix it does not get a new one.

Similarly, if the searcher views the same representation twice, i.e., the same summary twice, the heuristic-based framework only counts the representation once. The framework in effect, keeps a history of which representations have been viewed; it does not consider more detailed interaction. It is not possible to differentiate, for example, between a searcher seeking relevant information and a searcher checking what they have already examined, something that may account for them looking at the same representation twice.

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

6.2.3 Query Modification

In the matrix created by the Binary Voting Model, only the query terms and terms in representations viewed by the searcher will have a score greater than zero. The latter set of terms is potentially useful for query modification.

After every five relevance paths a new query is constructed. This number of paths was established through pilot testing and allows the model to gather sufficient implicit evidence from searcher interaction. It is possible for a relevance path to contain only one representation. Therefore, for the searcher to follow five paths they need only view five unique document representations.

To compute the new query the framework calculates the average score for each term across all documents (i.e., down each column in the document × term matrix). This gives an average score for each term in the vocabulary. The terms are then ranked by their average score. A high average score implies the term has appeared in many viewed representations and/or in those with high indicative weights across the documents viewed. The top six ranked terms are used modify the query. This modification can occur in two ways: *query expansion* and *query replacement*.

Query expansion – The top six terms chosen by the Binary Voting Model are appended to the original terms chosen by the searcher.

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

Example 6.1 (continued)

If for each term in the ten word vocabulary I average down all rows in the matrix the final weights for each of the ten terms in the vocabulary will now be:

$$\begin{array}{r|cccccccccc|}
 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 & t_{10} \\
Q_0 & 0 & 0 & 0 & 0 & .5 & 0 & 0 & 0 & .5 & 0 \\
D_{10} & .3 & .2 & .5 & .1 & .3 & .2 & 0 & 0 & .4 & .3 \\
D_5 & 0 & 0 & .1 & 0 & .1 & 0 & 0 & 0 & 0 & 0 \\
\hline
\text{average} & .1 & .07 & .2 & .03 & .3 & .07 & 0 & 0 & .3 & .1
\end{array}$$

The rank order of these terms after computation of these weights is t_5, t_9, t_3, t_{10}, t_1, t_2, t_6, t_4, t_7, t_8. Terms t_5 and t_9 are query terms and since the representations used in this model typically contain query terms I would expect these to be high ranked. In query expansion the top six terms would be added to the original query. In query replacement, the terms t_3, t_{10}, t_1 and t_2 would be appended to t_5 and t_9 to form a new search query.

The terms with the highest scores are those that are present most often in the information viewed by the searcher. The Binary Voting Model assumes that the non-query terms from these can be useful to represent the interests of the searcher. That is, the model considers terms with high weights are important and useful for query modification. The Binary Voting Model generates lists of terms at different temporal locations. The framework described in this chapter uses changes in the ordering of these term lists to estimate potential changes in the topic of the search. In the next section this process is described.

6.3 Information Need Tracking

The framework uses a history of recent interaction to predict changes in the information need of the searcher and make search decisions that may be useful in their search. This history provides insight into the recent interests of the searcher, and by comparing this with previous histories it can be used to track possible changes in the topic of the search. Selecting the most appropriate form of support depends on estimating the extent to which the need changes during a search; the smaller the change, the less radical the support offered. Tailoring the support in this way allows the interface to work in concert with the searcher. The degree of change between successive term lists (formed every five paths) provides evidence to approximate the degree of change in a searcher's information need.

In the set of most-relevant retrieved documents the vocabulary is static, so the framework can gauge the potential level of change in the information need by comparing the change in the term ordering from the term list at step *m* (i.e., L_m) and the list at the subsequent step *m*+1 (i.e., L_{m+1}). The term lists contain all terms in the vocabulary, ranked based on the weights assigned by the Binary Voting Model. As the vocabulary is static, the *terms* in the list will not change, only their order. So, by comparing L_m against L_{m+1} based on some operator $\bigcirc$ the framework can compute the degree of change between the lists and predict possible changes in the information need. This can be shown formally as:

$$\Delta\psi = (L_m) \bigcirc (L_{m+1}) \tag{6.2}$$

Where ψ is the system's view of the searcher's information need and $\bigcirc$ computes the difference between two lists of terms.

The *Spearman rank-order correlation coefficient* is used in this framework as the operator $\bigcirc$. This coefficient tests for the degree of similarity between two lists of rankings. The Spearman rank-order correlation coefficient is non-parametric, so rankings, not the actual term scores, are used. There are two lists of terms, L_m and L_{m+1}, created by the Binary Voting Model at two successive points in time. The first list is ordered by average term score; the second list contains the terms in the same order but updates the rankings (i.e., a new ranking is assigned, but there is *no sorting*). This is shown in Figure 6.4.

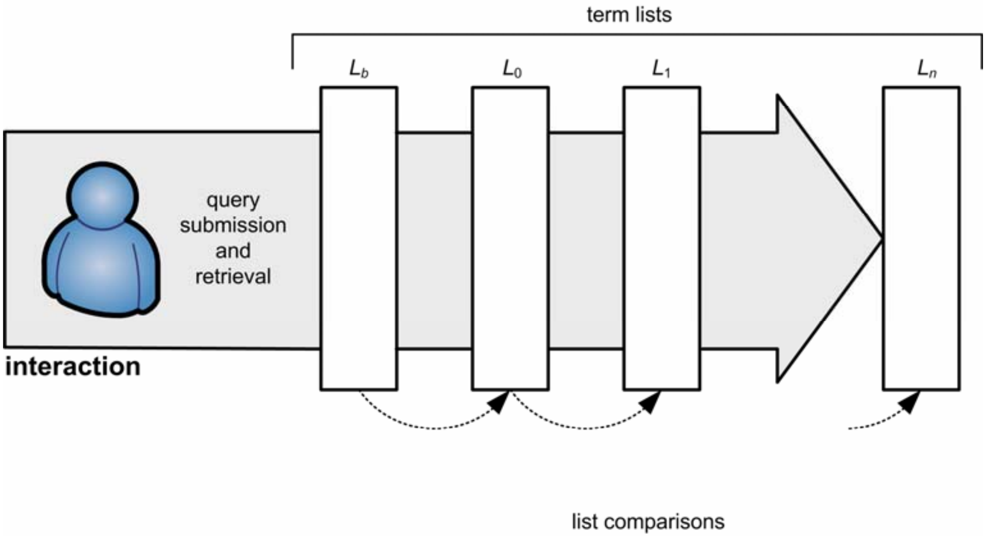


Figure 6.4. Changes in rank order of terms in consecutive term lists.

The Spearman rank-order correlation coefficient returns values between –1 and 1, where 1 is *perfect positive correlation* (the lists are exactly the same), –1 is *perfect negative correlation* (the lists are the complete opposite i.e., 1,2,3,4,5,6,7,8 vs. 8,7,6,5,4,3,2,1) and any value in-between is reflective of their relation to these extreme values. A correlation of 0 implies *zero* (or no) *correlation* between the two lists. Using the Spearman rank-order correlation coefficient $\Delta\psi$ is calculated as follows:

$$\Delta\psi = \frac{\sum_{i=1}^{n} r\left(L_{m_i}\right) r\left(L_{(m+1)_i}\right) - n\left(\frac{(n+1)}{2}\right)^2}{\left(\sum_{i=1}^{n} r\left(L_{m_i}\right)^2 - n\left(\frac{(n+1)}{2}\right)^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n} r\left(L_{m+1_i}\right)^2 - n\left(\frac{(n+1)}{2}\right)^2\right)^{\frac{1}{2}}} \qquad (6.3)$$

I assume that L_m and L_{m+1} are both ranked lists of terms, $r(.)$ is the rank of a term from one of the lists and n is the total number of terms. Ties are handled in the standard statistical way, by summing the rank of all tied elements and dividing this sum by the number of elements, effectively taking the average rank for each group of ties.

All terms in the original vocabulary [16] are ranked based on the weights derived from the Binary Voting Model, and averaged across all viewed documents. These terms are present in both lists (L_m and L_{m+1}) but potentially in a different order, depending on the representations viewed by the searcher. There is a high level of redundancy in each list as the lower ranking terms that never appear in a viewed representation experience only slight changes in their ranking between iterations. To counter this problem only the top 100 terms are used. These are the most liable to change and hence most likely to reflect any change in the information need. As the number of terms increases (i.e., greater than 100), redundancy in the term list also increases and the predicted level of change becomes more conservative. In contrast, as the number drops (i.e., less than 100) the likelihood of change increases, making the prediction more dramatic.

Term lists are compared every time a new query is created (i.e., every five relevance paths). To compute the correlation coefficient both lists must contain the same terms and the same number of terms. Therefore, in practice the first 100 terms plus β are used. Beta (β) is the number of terms that have left or joined the top 100 terms between L_m and L_{m+1}. For terms *joining* the top 100, these terms are sorted based on their original (L_m) ranks and assign them ranks (in L_m) in the range [101,101 + β]. The same procedure is used for terms that are *leaving* the top 100, except these terms are ranked based on their new (L_{m+1}) ranks (Figure 6.5).

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

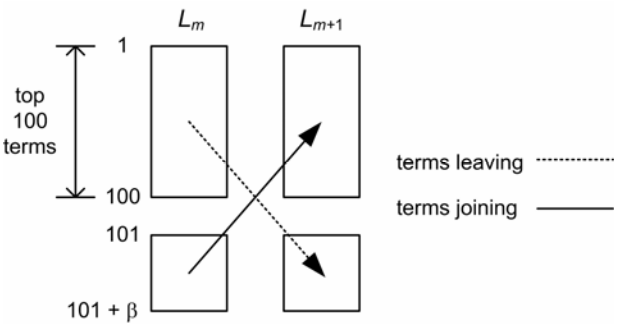


Figure 6.5. Terms leaving and joining the top 100 terms.

The Spearman coefficient is in the range [–1, 1], where a result closer to –1 means the term lists are dissimilar with respect to their rank ordering. Boundaries were chosen for the Spearman coefficient that allowed the need tracking component to choose retrieval strategies. These boundaries are shown in Figure 6.6.

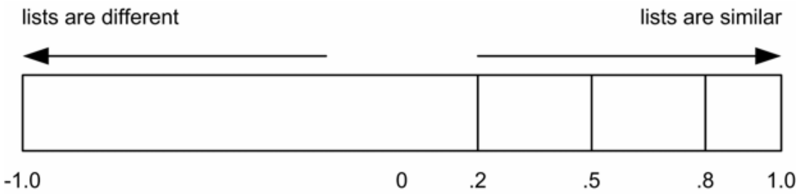


Figure 6.6. Decision boundaries of Spearman coefficient for retrieval strategy selection.

As the coefficient gets closer to one, the similarity between the two query lists increases and based on the coefficient value the framework decides how to use the new list of terms. Four retrieval strategies were implemented:

Re-searching – If the coefficient value indicates that the two term lists are substantially different with respect to rank ordering, I take this to reflect a large change in ψ (the system's formulation of the information need). In this case, a search system implementing the framework will re-search the document collection to retrieve a new set of documents. Coefficient values of less than 0.2 are taken to indicate a large change in the term lists.

Reordering documents – A result in the range [0.2, 0.5) indicates a weak correlation between the two lists and consequently a less substantial change in ψ. Here an implementing search system will use the new query (i.e., the six top ranked terms) to reorder the most-relevant retrieved documents. The document list is reordered using best-match *tf.idf* scoring with the revised query. The vocabulary list remains unchanged after this action.

Reordering Top-Ranking Sentences – Coefficients in the range [0.5, 0.8) indicate a strong correlation between the two term lists and hence a small change in the system's estimation of information needs. In this case the framework uses the new query to re-rank the list of Top-Ranking Sentences. The sentences are the most granular elements presented to the searcher and are therefore most suited to reflect minor changes in ψ. The Top-Ranking Sentences are reordered based on the term-occurrence of each of terms in the new query.

No action – The previous strategies provide an updated view of the retrieved documents based on the current ψ. For differences between 0.8 and 1 (inclusive), the need is assumed to have not changed sufficiently to warrant action.

All numerical bounds are experimental, chosen during pilot testing of the framework. This involved testing an experimental system that implemented the framework interactively with different levels of search topic change (i.e., viewing information on one topic then looking at another). As I viewed information at the results interface the actual value of the Spearman correlation coefficient and boundaries assigned were displayed graphically in a small window in the results interface (Figure 6.7).

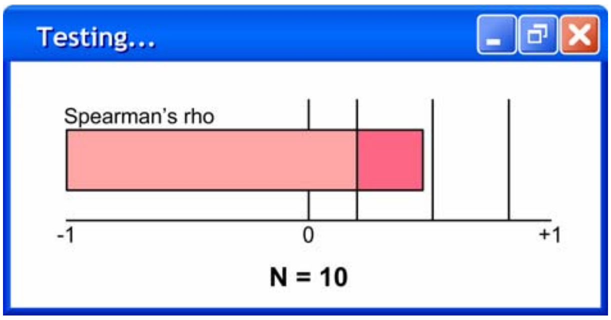


Figure 6.7. Monitoring Spearman rank correlation coefficient during pilot testing.

The lines representing the boundaries could be dragged to different values. Over time, and a variety of search topics, the boundaries were placed in a location that resulted in a high proportion of system decisions being deemed appropriate. This was subjective but was tested in Pilot Test 1 described in Chapter Nine, Section 9.2.1.

6.4 Summary

In this chapter a heuristic-based implicit feedback framework for estimating information needs during a search has been described. The approach uses a Binary Voting Model to create a modified query and a decision metric based on Spearman's rank order correlation coefficient to predict changes in the topic of the search and make new search decisions. The evaluation of this framework using human subjects is described in Chapter Nine (Pilot Test 1). In the next chapter a probabilistic framework for selecting additional terms and retrieval strategies is presented. This framework is potentially more robust than that presented in this chapter and formalises some of the heuristics used.

Chapter 7

Probabilistic Framework

7.1 Introduction

In this chapter a second implicit feedback framework is proposed to estimate current information needs and changes in these information needs during a search session. This framework uses interaction with document representations and relevance paths in the same way as the heuristic-based framework described in Chapter Six. Also, in a similar way to that framework, the techniques presented create modified query statements based on implicit evidence and includes a component to predict when, and by how much, information needs have changed. The approach to select terms is probabilistic and uses *Jeffrey's rule of conditioning* (Jeffrey, 1983) to revise the probability of term relevance in light of evidence gathered from searcher interaction; this is called the *Jeffrey's Conditioning Model*. Jeffrey's conditioning captures the uncertain nature of implicit evidence, and is used since even after the 'passage of experience' the model is still uncertain about the relevance of a term. The approach used for this revision is based on that proposed by Van Rijsbergen (1992). The information need tracking component to estimate need change uses Pearson's correlation coefficient and the statistical significance of this coefficient as a decision metric. I describe the information need detection and need tracking components in Sections 7.2 and 7.3 respectively.

7.2 Information Need Detection

This component uses interaction with representations of top-ranked retrieved documents and relevance paths to predict the interests of searchers. The Binary Voting Model (Chapter Six) used a set of pre-defined heuristic weights for the indicativity of a relevance path's constituent representations. The information need detection component in the probabilistic model replaces this with a measure to describe the value, or worth, of the evidence in a document representation. It combines a confidence measure that uses the relative position of representations in a relevance path with a measure of indicativity based on the concepts in a representation. Unlike the Binary Voting Model, the probabilistic model uses relevance paths directly in the revision of term probabilities. In this section I describe each measure, and how the probabilistic model weights terms for query modification.

7.2.1 Path Weighting

As described earlier in this thesis, Campbell and Van Rijsbergen (1996) produced an extension of the probabilistic model of retrieval that incorporates an 'ageing' component to term weighting. The component incorporates *when* documents containing the terms were assessed relevant. Searchers follow a path through the document space and terms in documents assessed relevant at an early stage in the search receive a lower weight than terms in recently viewed documents.

In content-rich search interfaces searchers traverse relevance paths between document *representations*. Unlike the work of Campbell and Van Rijsbergen, the representations that comprise the path are smaller than documents, the paths are generally short (i.e., no more than six representations) and the most recent document representation is not necessarily the most relevant. The term selection model described in this section assigns an exponentially increasing relevance profile to aged relevance.

The assumption made by this part of the framework 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 as information need* the model is cautious, and hence less confident about the value of this evidence. This confidence, c, is assigned *from the start of the path* to each representation i using:

$$c_i = \frac{1}{2^i}, \text{ where } i \geq 1 \tag{7.1}$$

However, Equation 7.1 is asymptotic, and therefore the values of c_i do not sum to one. For this to be used in the information need detection component (which is based on probabilistic principles) the sum of all c_i values must be one. The value of c_i is normalised and the confidence c for each representation i in a path of length N is computed using:

$$c_i = \left(\frac{1}{2^i} + \frac{1}{N.2^N} \right), \text{ where } \sum_{i=1}^{N} c_i = 1 \text{ and } i \in \{1,2,...,N\} \quad (7.2)$$

Document representations are weighted based on the confidence in their contribution; those near the start of paths are assumed to drive interaction along the path, and are given more weight than those at the end. The representations are weighted based on their position in the relevance path and for the information they contain. A good document representation should be indicative of the source document and in the next section I describe how indicativity is determined.

7.2.2 Indicativity and Quality of Evidence

In the previous section I described the confidence in the relevance of representations based on their position in the relevance path. The profile assumes that subsequent steps in the relevance path are half as relevant; this is perhaps too severe. The quality of evidence in a representation, or its *indicative worth*, can also affect how confident the framework is about the value of its content. In the Binary Voting Model I used heuristics based on the *typical* length of document representations to measure indicativity. However, titles and Top-Ranking Sentences, which may be very indicative of document content, are short and will have low indicativity scores if their typical length is the attribute used to score them.

In this framework I used the non-stopword terms, or *concepts*, in a representation as a measure of indicativity rather than representation length. The weight of a term *t* in document *d* is calculated using its *normalised term frequency* (Harman, 1986), and normalised so that the sum of all weights in a document is one. The larger this value, the more often it occurs in the document, and the more representative of document content that term can be seen to be. To compute the indicativity index *I* for a representation *r* the weights of a term in a document $w_{t,d}$ were summed for all *unique* terms in *r* such that:

$$I_r = \sum_{t \in r} w_{t,d} \quad (7.3)$$

The I_r ranges between zero and one, is never zero, and is one only if the representation contains every unique term in the document. The indicativity measure is only incremented if there is a match between the unique terms in the document and those in the representation. [17] If one assumes that a document *d* contains the set of terms s and representation of that

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

document *r* contains the terms $\{t_1, t_4, t_{10}\}$. Since terms t_1, t_4 and t_{10} occur in both the representation and the source document, the indicativity index of *r* would be:

$$I_r = w_{t_1,d} + w_{t_4,d} + w_{t_{10},d}, \text{ where } 0 < I_r \leq 1 \tag{7.4}$$

Relevance paths will contain representations of varying quality. The indicativity of a representation can also be seen as measure of the *quality* of the evidence provided by that representation. The indicativity of a representation is multiplied with the confidence associated with that particular step in the relevance path (from Equation 7.2) to compute the *value* of the evidence. Using these measures helps ensure that the worthwhile representations in each relevance path contribute most to the selection of potentially useful query modification terms. In the next section the approach used to select such terms is described.

7.2.3 Term Weighting

The probabilistic model assumes the existence of a *term space T*, a mutually exclusive set of all (non-stemmed, non-stopword) terms in the set of top-ranked retrieved documents. Each term in *T* is independent and has an associated frequency in the top documents. The normalised term frequency of each term in *T* is used to estimate its probability. The probability that a term *t* is relevant based on a probability distribution *P* over *T* as:

$$P(t) = \frac{ntf(t)}{\sum_{t \in T} ntf(t)} \text{ where } ntf(t) = \frac{\log_2(tf(t)+1)}{\log_2|T|} \tag{7.5}$$

where *ntf* (*t*) is the *normalised term frequency* (Harman, 1986) of term *t* in the term space *T*.

To update this probability based on new evidence gathered from interaction the component uses *Jeffrey’s rule of conditioning* (Jeffrey, 1983), applied at the end of each relevance path. Jeffrey’s rule is a generalisation of Bayesian belief revision. The basis for Bayes’ Theorem (Bayes, 1763) is:

$$P(H \mid e) \propto P(e \mid H)\, P(H) \tag{7.6}$$

or

$$P(H \mid e) = \frac{P(e \mid H)P(H)}{P(e)} \text{ since } \sum_H P(H \mid e) = 1 \tag{7.7}$$

Where *H* is a hypothesis that is supported (or refuted) by some evidence *e* and *P*(*H* | *e*) is interpreted as the probability that this hypothesis is true given *e*. Bayes’ Theorem is regarded as a form of belief revision since the component probabilities – $P(e \mid H)$, $P(H)$ and $P(e)$ – are associated with propositions (or events) prior to e being observed and $P(H \mid e)$ is the probability after observation. This can be problematic as Bayesian belief revision requires the evidence e to be certain when it is observed and cannot be disputed (Van Rijsbergen, 1992).

In IR the Bayesian approach is typically used for computing the probability of relevance R given a document and a query through $P(R \mid x)$ where x is some representation of the document assumed to be certain. This means that the description x is assumed to be true of, or in, the document. This may not always be appropriate as the document description may be uncertain at the time of observation. So, for example, a component of a document representation, x_i, might only apply to the source document with a certain probability (Maron and Kuhns, 1960). Since x_i is not certain, Bayes' Theorem cannot compute $P(R \mid x_i)$.

The problem of how to revise a probability measure in light of uncertain evidence or observation was covered by Richard Jeffrey in his book 'The Logic of Decision' (1983). His approach is best described by an example taken from his book.

Imagine a person inspects a piece of cloth by candlelight and thinks that it is green, although it might be blue, or even violet. If G, B and V are the propositions involved then the outcome of the observation might be that the degrees of belief in G, B or V are .70, .25 and .05, whereas *before* observation the degrees of belief were .30, .30 and .40. Represented formally this is:

$$\begin{array}{lll} P(G) = .30, & P(B) = .30, & P(V) = .40 \\ P'(G) = .70, & P'(B) = .25, & P'(V) = .05 \end{array}$$

Here P is a measure of the degree of belief before observation, and P' the measure after observation. The 'passage of experience' has led to P being revised to P', as in $P'(x) = P(x \mid e)$ where e is a proposition, but Jeffrey claims that it is not always possible to express the passage of experience as a proposition. Pearl (1988) used a Bayesian net formalism to make Bayesian conditionalisation appropriate, which was implemented by Turtle and Croft (1992). However, Van Rijsbergen (1992) suggests that Pearl's presupposition of virtual evidence leads to infinite regress (i.e., always being dependent on prior evidence) and that it is better to assume from the beginning that the passage of experience leads to a direct revision of the probability functions.

Given that the person has changed their degree of belief in some propositions G, B, and V as shown above, these changes must be propagated over the rest of the structure of their beliefs. For example, suppose saleability A of the cloth depends on the colour inspection in the following way:

$$P(A \mid G) = .40, \qquad P(A \mid B) = .40, \qquad P(A \mid V) = .80$$

Prior to inspection:

$$P(A) = P(A \mid G)\, P(G) + P(A \mid B)\, P(B) + P(A \mid V)\, P(V)$$
$$= .40 \times .30 + .40 \times .30 + .80 \times .40 = .56$$

After inspection Jeffrey proposes:

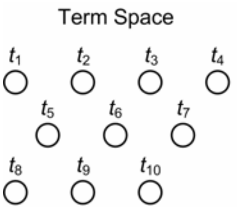
$$P'(A) = P(A \mid G)\, P'(G) + P(A \mid B)\, P'(B) + P(A \mid V)\, P'(V)$$
$$= .40 \times .70 + .40 \times .25 + .80 \times .05 = .0485$$

This is *Jeffrey's rule of conditioning*. This differs from Bayesian conditioning which would use $P'(G) = 1$, or $P'(B) = 1$, or $P'(V) = 1$ and so revise $P(A)$ to $P'(A) = P(A \mid X)$ when $X = G$, B, or V. Bayesian conditioning can therefore be viewed as a special case of Jeffrey's conditioning.

In the context of the work presented in this chapter, a relevance path p is considered as a new source of evidence to update the probability P to P'. I now describe the term weighting approach through an example.

Example 7.1: Simple Updating

Assume the existence of a term space containing 10 terms with the initial values as shown in Figure 7.1. The initial query (Q_0) contains the terms t_5 and t_9. The term space is based on the searcher's query and is therefore created dynamically from the retrieved set of documents; since this set is topically relevant, and the query terms are weighted highly in the initial term space.



	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
$tf(t)$	1	1	2	1	5	2	1	1	5	1
$P(t)$	.07	.07	.11	.07	.18	.11	.07	.07	.18	.07

Figure 7.1. Initial term space frequencies and probabilities for Example 7.1.

The initial $P(t)$ of terms in the term space is determined based on Equation 7.5. This value is normalised to ensure the sum of all probabilities is one. However, it is necessary to revise these probabilities in light of new evidence. In the next section I describe this process.

7.2.4 Probability Revision

The top-ranked documents from which the term space is derived contain a number of document representations. These representations are presented to searchers at the content-rich search interfaces.

The viewing of a representation p_i creates new evidence for the terms in that representation. Let us consider the property of relevance and let us consider the effect of observing an index term t, which is either present ($t = 1$) or absent ($t = 0$). Then:

$$P'(t) = P(p_i \mid t = 1)\, P'(t = 1) + P(p_i \mid t = 0)\, P'(t = 0) \tag{7.8}$$

This conditional probability may also be estimated through the standard Bayesian inversion using the following formula:

$$P'(t) = \left[P(t = 1 \mid p_i) \frac{P'(t = 1)}{P(t = 1)} + P(t = 0 \mid p_i) \frac{P'(t = 0)}{P(t = 0)} \right] . P(t) \tag{7.9}$$

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 $P(t = 0)$ the probability of not observing t. The prior estimate $P(t = 1)$ is given by collection statistics using Equation 7.5. The probabilities $P'(t = 1)$ and $P(t = 1 \mid p_i)$ are computed in the same way as $P(t = 1)$ (i.e., with Equation 7.5) with one difference in each case; rather than using the frequency of term t in the top documents, $P'(t = 1)$ uses the frequency of t in the whole relevance path and $P(t = 1 \mid p_i)$ uses the frequency of t in the representation p_i. The updated probability $P'(t)$ reflects the 'passage of experience' and is similar to that described by Van Rijsbergen (1992).

A relevance path contains a number of representations. The probabilities are updated after the traversal of a relevance path. The length of a relevance path ranges between one and six steps. I denote this length using *N*. When this length is greater than one 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*:

$$P_N(t) = \sum_{i=1}^{N-1} c_i.I_i.\left[\left(P_i(t=1\mid p_i)\frac{P_{i+1}(t=1)}{P_i(t=1)} + P_i(t=0\mid p_i)\frac{P_{i+1}(t=0)}{P_i(t=0)}\right) .\ P_i(t)\right] \qquad (7.11)$$

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.

In Bayesian belief revision the order of conditioning is irrelevant. However, in Jeffrey's conditioning this is not the case and in general the order the evidence is presented does matter. Therefore the order in which a searcher traverses the relevance path also matters. The content-rich search interface described in Chapter Five restricts the order in which relevance paths can be traversed. However, where a top-ranking sentence appears in the path i.e., in the list of sentences at the start of the path or as a summary sentence, can affect how much that sentence contributes to the selection of terms. A top-ranking sentence contributes more to the selection of new query terms than the same sentence appearing as a 'summary sentence' later in the relevance path; the framework does not weight the same evidence to the same extent twice. This seems reasonable as it was the top-ranking sentence that encouraged the searcher to initiate the traversal of relevance path. The term selection component is uncertain whether further traversal is explorative (i.e., to see what information is available) or verificative (i.e., to verify initial perceptions of relevance) and assigns a reduced weight to representations that fall later in the relevance path to represent this uncertainty.

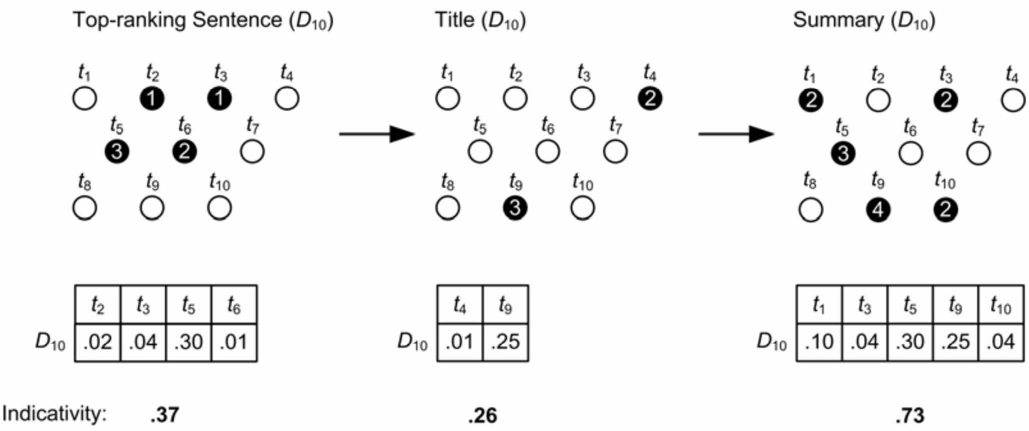
As will be described later in this section, the actual revision of the probabilities will occur after each path. Once learned, the probabilities of relevance remain stable until the next revision (i.e., the next relevance path). Only terms in *T* that appear in the relevance path will have their probabilities revised directly. [18] This order of updating is sequential in nature; that is, relevance paths provide pieces of evidence that are taken sequentially. There is scope for this to allow a searcher to change their mind about the strength of evidence or reverse each revision step. However, this is not implemented in this thesis as revisions occur based on implicit evidence that the searcher may not be aware is being captured or may not be involved

[18] Based on the new evidence probabilities are redistributed to make the sum one.

directly in its provision; they may therefore not know *when* it is appropriate reverse revisions or change the strength of evidence.

Example 7.1: Simple Updating (continued)

When a searcher follows a relevance path, the term selection model updates the weights in the term space after each path. Figure 7.2 shows how the term weights are updated as a path. In this example one can assume the searcher has expressed an interest in a top-ranking sentence from document D_{10}, then the title, and finally the summary. The numbers inside the ball for each term (e.g., ❶) indicate the term frequency in that representation.


Top-ranking Sentence (D_{10})
t_1 ○ t_2 ❶ t_3 ❶ t_4 ○
t_5 ❸ t_6 ❷ t_7 ○
t_8 ○ t_9 ○ t_{10} ○
Title (D_{10})
t_1 ○ t_2 ○ t_3 ○ t_4 ❷
t_5 ○ t_6 ○ t_7 ○
t_8 ○ t_9 ❸ t_{10} ○
Summary (D_{10})
t_1 ❷ t_2 ○ t_3 ❷ t_4 ○
t_5 ❸ t_6 ○ t_7 ○
t_8 ○ t_9 ❹ t_{10} ❷
Indicativity:
.37
.26
.73

	t_2	t_3	t_5	t_6
D_{10}	.02	.04	.30	.01

	t_4	t_9
D_{10}	.01	.25

	t_1	t_3	t_5	t_9	t_{10}
D_{10}	.10	.04	.30	.25	.04

Figure 7.2. Indicativity at each step in relevance path in Example 7.1.

In this framework the quality of a representation is a measure of its indicativity; how well it represents the source document. A decreasing ostensive relevance profile is applied across the relevance path normalised by the indicativity values shown above in Figure 7.2. This contrasts with the *decreasing* ostensive relevance profile described by Campbell (2000), which assumes that a representation directly following the current representation is half as relevant (shown as a dotted line in Figure 7.3). This seems simplistic, too severe and a normalisation of this profile that includes the quality of the representations in the path is perhaps more fitting. To create such a measure, I multiply the ostensive weight of each step in the relevance path (from Equation 7.2) by its indicativity to form the normalised ostensive relevance (shown by the solid line in Figure 7.3).

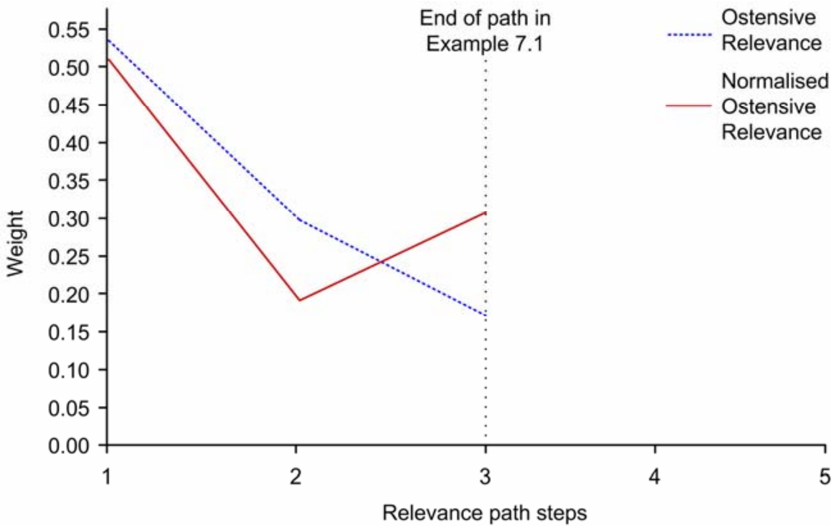
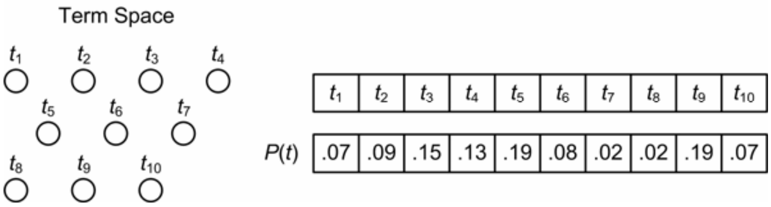


Figure 7.3. Ostensive Relevance and Normalised Ostensive Relevance profiles in Example 7.1.

The weights are stored temporarily in variables until the complete path is traversed. The computation of the revised probabilities is dependent on knowing the length of the relevance path. For this reason it is necessary to wait until the complete path has been traversed before calculating the new set of probabilities. The temporary variables are reset after each relevance path.

The weights of all terms bar t_7 and t_8 are directly updated. Since t_7 and t_8 do not appear in any viewed representations their weights are updated indirectly to ensure the sum of $P(t)$ is one. This can be interpreted as a form of *negative* relevance feedback as the weights of these terms will decrease. Figure 7.4 shows the final state of the term space after all revisions.



	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
$P(t)$	.07	.09	.15	.13	.19	.08	.02	.02	.19	.07

Figure 7.4. Term space in Example 7.1 after relevance path.

The final term ordering, based on the $P(t)$ for each term is t_5, t_9, t_3, t_4, t_2, t_6, t_1, t_{10}, t_7 and t_8 [19]. The terms with the highest scores will be chosen to replace or expand the searcher's original

[19] This is in contrast to the term ordering t_5, t_9, t_3, t_{10}, t_1, t_2, t_6, t_4, t_7, t_8 that arises when the same evidence is presented to the Binary Voting Model in Chapter Six, Example 6.1. The differences in ranking arise because the Jeffrey's Conditioning Model does not simply use term presence or absence and is more sensitive to the *frequency* of terms in document representations.

query. Terms that do not appear in viewed representations have the probability that they are relevant revised downwards. Since t_7 and t_8 do not appear in any of the viewed representations their $P(t)$ is revised indirectly downwards. In contrast, t_3 appears in two of the representations in the relevance path and its $P(t)$ is revised upwards from 0.11 to 0.15.

In this section an approach has been described for estimating the information needs of searchers at a specific point during their search. However, to operate effectively the framework must also be able to identify when a search has changed (i.e., moved from one topic to another). Information needs are dynamic and can change in dramatic or gradual ways as the searcher views information (Bruce, 1994; Robins, 1997). In the next section I describe an information need tracking component similar to that described in Chapter Six that predicts the level of change in a searcher's information need and makes search decisions to help them search effectively.

7.3 Information Need Tracking

In Chapter Six an approach for estimating potential changes in the information needs of searchers was described. This change is based on the differences between the ordered lists of terms, created by the term selection model during the search. Since the order of the terms is affected by the information the searcher expresses an interest in, changes in the term order can be used to track when and by how much a search has changed.

The heuristic approach in Chapter Six uses the Spearman rank order correlation coefficient to provide an estimation of search topic change. The approach worked well, and seemed to choose retrieval strategies that were appropriate. However, in related work (White and Jose, 2004) I have shown that another correlation coefficient, the Pearson product moment (r) also concords with searcher opinion. In this section a similar, but more robust approach is described that uses Pearson's r to estimate search topic change. The statistical significance of r is used to select new retrieval strategies based on the extent of the change and as a further measure of confidence in system decisions.

At every point in the search the contents of the term space are ranked in descending order based on the value of $P(t)$. The order of this term list may change when all $P(t)$ are revised. It is the level of difference between the two lists of terms at different temporal locations that can be used by the framework to predict search topic change.

The information need tracking component in Chapter Six compared consecutive term lists in such a way that the current list of terms was only compared against its immediate predecessor. This meant that the information need would have to change a lot between those iterations for it to be detected by the framework. Rather than watching for changes in information needs over the past few steps, the approach presented in this section looks for changes since the last searcher-defined query iteration. That is, since the searcher last modified their query and used it to generate a new set of search results. The approach uses the interaction that immediately follows the presentation of search results to generate a baseline term ordering (L_b). Each term list (L_i) from that point until the next *re-search* operation is compared to L_b. This is illustrated in Figure 7.5.

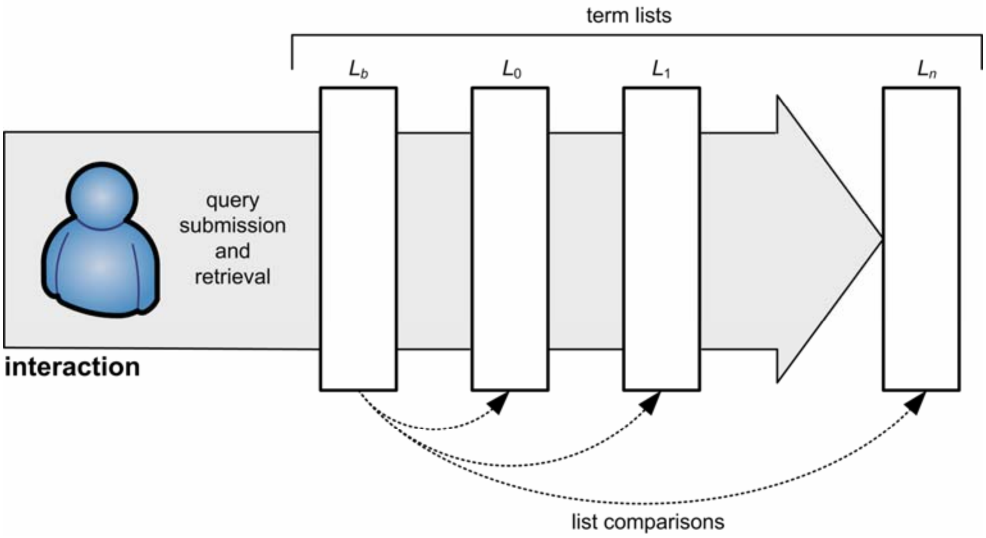


Figure 7.5. Comparing $P(t)$ of terms between term lists.

The approach described in Chapter Six looked for changes in only the top 100 terms in the list. A vocabulary constructed from the top-ranked documents would typically contain some 3000 unique words, depending on the nature of the retrieved documents. This introduces redundancy into the correlation coefficient calculations, as many of the low-ranked terms will experience only slight changes in their ranking during a search session. To address this problem I established a threshold and only used terms whose scores placed them in the top 100. Analysis of the system logs from my experiment found that this threshold was too low, as many terms outside the top 100 changed position also.

The approach proposed in this chapter does not use thresholds; instead it uses the changes in ordering of all terms in viewed representations between query iterations. If a term appears in a viewed representation it is considered an 'active' term and is used in the calculation of

Pearson's *r*. Terms which are 'inactive', i.e., do not appear in viewed representations, are ignored. I effectively only consider changes those terms whose probability is revised *directly* at some point during the search.

Using correlation coefficients is not the only method that can be used to detect the similarity of the two term lists. Measures of *association* (e.g., Cosine, Jaccard's, Dices), *distance* (e.g., Euclidean, L1(norm)) and *divergence* (e.g., Kullback-Liebler, Jensen-Shannon) can also be used. [20] The disadvantage of these methods is that statistical significance cannot be derived from them. Kupperman (1960) proposed an information statistic based on the χ^2-distribution that estimated the statistical significance of the divergence between two probability distributions using Kullback-Liebler divergence. However, the size and variability of the degrees of freedom (ranging from 0 to 3000) meant that the critical value of χ^2 and hence statistical significance of the information measure, had to be computed in real-time (i.e., it could not be looked up in statistical tables). For reasons of simplicity and reduced computational expense this statistic was not used in this framework. Pilot testing of this statistic for tracking search topic change showed that the statistic could be unreliable, choosing retrieval strategies that were at times inappropriate. In a related pilot study (White and Jose, 2004), the Kullback-Liebler divergence was shown to be a poor predictor of searcher assessments of search topic change. Since the Spearman rank order correlation coefficient (r_s) worked well in Pilot Test 1 (a user experiment described in Chapter Nine, Section 9.2.1) and there was no need to over-complicate this part of the framework. Pearson's *r* was used in place of r_s as it is parametric, hence based on the values of $P(t)$ not their rank order. In r_s information is lost in the conversion from measurements to rankings meaning that *r* is more sensitive to small changes in the distribution.

The t-distribution is used to compute the statistical significance of Pearson's *r* using the formula shown in Equation 7.10 with $N - 2$ degrees of freedom, where N is the number of pairs. The t-distribution is defined as the distribution of the random variable t which is (very loosely) the 'best' that can be calculated not knowing the standard deviation. The test of significance allows us to estimate the probability that there is a relationship between the two term lists.

Pre-defined thresholds are used in a similar way to that shown in Figure 6.6. These techniques are used in conjunction with tests of statistical significance that test the statistical validity of a claim that the difference between the term lists occurred by chance.

[20] For a summary of these measures see Lee (1999).

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{7.10}$$

The null hypothesis states that there is no relationship between the two lists (i.e., $r = 0$). [21] The alternative hypothesis states that there is a true relationship i.e., significantly similar or significantly dissimilar. The value of t and the number of degrees of freedom can be used to predict the statistical significance of the correlation between the two term lists.

The level of significance can be used to help the framework make decisions on which retrieval strategy to employ. The framework must choose one of four possible retrieval strategies (in decreasing order of severity): *re-search Web*, *reorder documents*, *reorder Top-Ranking Sentences* and *no action*. These strategies allow the retrieval system to recreate or restructure the retrieved information based on their most recent estimations of the information need. Firstly the framework obtains a value for r and chooses a strategy in a similar way to that proposed in Chapter Six, although with revised threshold values. [22] It then tests the significance of the correlation. If r is significant at $p < .05$ (two-tailed test) the retrieval strategy proceeds, if not (i.e., $p \geq .05$) then a new retrieval strategy, of lesser severity, is chosen. For example, if the initial decision is to *re-search Web* and:

$$N = 20 \qquad r = .254$$

$$t(18) = 1.114 \qquad \mathbf{p = .279}$$

Since $p > .10$ the decision would be modified to the more conservative option of *reorder documents*. The statistical significance of the difference therefore contributes to the decision about which retrieval strategy should be employed. This happens for all strategies except *no action*, which is not affected since it is the most conservative retrieval strategy. The bounds used by the framework to choose the retrieval strategy are illustrated in Figure 7.6.

[21] If the null hypothesis suggests that r is above or below zero then one must use Fisher's transformation to first make the distribution of r normal, then compute Z-scores and finally p values.

[22] An alternative would be to convert r to r^2 to determine the strength of the correlation. However, this does not consider the direction of the correlation, effectively regarding $-r$ as the same as $+r$. Since a negative correlation can be interpreted as an indication of larger difference the sign of the coefficient is important in this part of the framework.

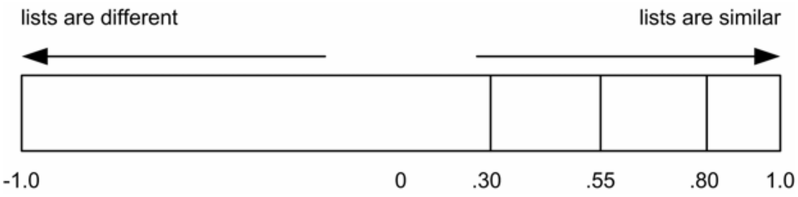
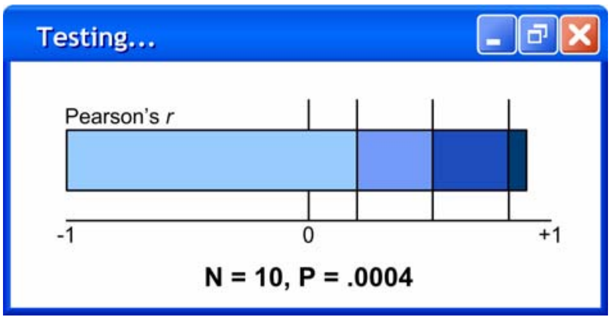


Figure 7.6. The decision boundaries of Pearson's *r*.

The boundaries were *re-search Web* [-1.0, .30), *reorder documents* [.30, .55), *reorder Top-Ranking Sentences* [.55, .80) and *no action* [.80, 1.0]. The boundaries differ slightly from those in the information need tracking component in the heuristic-based framework although the bounds were chosen through pilot testing in a similar way to Chapter Six. The differences between frameworks are attributable to the increased sensitivity of Pearson's *r* (which compares actual values) over Spearman's r_s (which uses rank ordering). For this reason the boundaries originally set to .20 and .50 in Figure 6.6 (Chapter Six) were increased to .30 and .55 respectively, to improve the reliability of the retrieval strategy selection.

To determine the boundaries I used a system that implemented the approach interactively with different levels of search topic change (i.e., looking at information on one topic then looking at another). As I viewed information at the results interface the decision of the system, the actual value of *r* and boundaries assigned were displayed graphically in a small window in the results interface (Figure 7.7).



N = 10, P = .0004

Figure 7.7. Monitoring Pearson's *r* during pilot testing.

The lines representing the boundaries could be dragged to different values. Over time, and a variety of search topics, the boundaries were placed in a location that resulted in a high proportion of appropriate system decisions. These assessments were subjective and the effectiveness of the information need tracking component was further tested in the user experiment described in Part IV.

7.4 Summary

In this chapter a framework for estimating and tracking changes in information needs is presented. The approach uses Jeffrey's rule of conditioning to revise the probability of relevance for all terms in light of new (uncertain) evidence. The most relevant terms are assumed to best represent information needs. In this chapter I have also proposed an approach using Pearson's product moment correlation coefficient to estimate changes in information needs within search sessions and use this estimation to make new search decisions.

In the next chapter I describe a novel simulation-based evaluation that tests the Binary Voting Model described in Chapter Six and the Jeffrey's Conditioning Model described in this chapter. The evaluation compares the performance of these models and other term selection baselines. Later in this thesis I evaluate the implicit feedback frameworks with human subjects; the experimental methodology employed and the results of this experiment are presented in Part IV.

Chapter 8

Benchmarking Implicit Feedback Models

8.1 Introduction

So far in Part III implicit feedback frameworks that use interaction with content-rich interfaces to approximate information needs have been described. In this chapter I introduce a simulation-based evaluation methodology to evaluate the search effectiveness of a variety of term selection (implicit feedback) models independent of searchers. I propose a Simulated IMPLicit Evaluation approach called 'SIMPLE' that simulates interaction with the search interface described in Chapter Five. The models evaluated include, among others, the information need detection components from the heuristic-based and probabilistic frameworks described in Chapters Six and Seven.

There is no standard way to evaluate term selection models that require a lot of searcher interaction with results interfaces. Typically, the only interaction modelled in standard IR experimentation is the provision of relevance feedback through marking relevant documents (Buckley *et al.*, 1994); this is relatively simplistic. The interaction required to provide feedback for the models described in Chapters Six and Seven is complex and a methodology that simulates such interaction is required. 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 less time consuming and costly than experiments with human subjects, allow environmental and situational variables to be more strictly controlled and complex searcher interactions to be modelled. The SIMPLE approach allows the comparison and fine-tuning of the term selection models before they are employed in an operational IR system. The best performing model is selected and used as the term selection model in the implicit feedback framework tested further in the user experiment described in Part IV. I call components that select additional terms for query modification based on implicit feedback from searchers *implicit feedback models*.

In this chapter I evaluate only the information need *detection* part of the frameworks, not the information need *tracking* component. The information need tracking components are tested through experimentation with human subjects (described in Part IV), not simulations since it may be difficult to assess their effectiveness objectively. Six implicit feedback models are tested using the searcher simulations described in this chapter. These include the Binary Voting and Jeffrey's Conditioning Models from the frameworks described in Chapters Six and Seven and other baseline models. In the next section I describe these baselines.

8.2 Baseline Implicit Feedback Models

In this study I investigate a variety of different methods of relevance feedback weighting based on implicit evidence. The implicit feedback models presented use different methods of handling this implicit evidence and revising their beliefs about searcher needs in light of it. The simulations present the model with evidence in the form of document representations, relevance paths that join representations and the full-text of documents. The study compares the models' ability to 'learn' [23] relevance and create more effective search queries. The performance of the Binary Voting Model (Chapter Six), the Jeffrey's Conditioning Model (Chapter Seven) and four baselines that use a variety of methods to choose additional query terms are compared. Three of the baseline models use the popular *wpq* query expansion method (Robertson, 1990) and one model selects terms randomly. These models are described in subsequent sections.

8.2.1 WPQ-Based Models

The *wpq* method (Robertson, 1990) has been shown to be effective and produce effective query enhancements 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, N = the number of documents in the collection.

$$wpq_t = \log \frac{r_t/(R-r_t)}{(n_t-r_t)/(N-n_t-R+r_t)} \cdot \left(\frac{r_t}{R} - \frac{n_t-r_t}{N-R} \right) \tag{8.1}$$

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

In the models described in this chapter, 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 non-relevant documents. As the values of r_t and R change during searcher interaction, the *wpq*-generated term weights also change. However, there is no retained memory of these term weights between iterations, and wpq_t is recomputed after each iteration. The *wpq* approaches learn what search results are relevant but do not directly 'remember' the weights assigned to *terms*. For example, a model based on *wpq* may be aware which documents have been explicitly marked but since these documents may change the term weights will have to be re-computed from zero at every iteration. In contrast, the Jeffrey's Conditioning and Binary Voting Models, store and revise term weights for the entire search session. At any point the term space, or vocabulary, stores the term weights for all potential query modification terms.

8.2.1.1 WPQ Document Model

The *wpq document model* uses the full-text of documents, rather than granular representations or paths that link them. The *wpq* formula is applied to each document and expansion terms chosen from it. In Equation 8.1 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 relevance feedback model, choosing one relevant document per iteration. This is a realistic model since implicit feedback is typically gathered sequentially (i.e., one relevance indication after another) and was included in the study to investigate the effects of using whole documents for such feedback.

8.2.1.2 WPQ Path Model

In the *wpq path model* the terms from each complete relevance path are pooled together and ranked based on their *wpq* score. The values 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, n_t = the number of generated paths that contain the term t are used in for the variable values in Equation 8.1. 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.

8.2.1.3 WPQ Ostensive Profile Model

The *wpq ostensive* [24] *profile 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 most recently viewed step in the path. The *wpq* weights are normalised 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, 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.

8.2.2 Random Term Selection Model

The random term selection model assigns a random score between zero and one 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 or search results this model was also expected to experience no learning.

In the next section I describe the simulated-based evaluation methodology used to test each of the six implicit feedback models.

8.3 Simulation-Based Evaluation Methodology

There has been no precedent set on how to evaluate implicit feedback models. In this study a simulation-based evaluation methodology is used to simulate interaction with the style of results interface described in Chapter Five, to benchmark such models and choose the best performing model to be further tested in the user experiment described in Part IV. This simulation-based study is therefore a formative evaluation of the implicit feedback models, in which only the best model is chosen for further experimentation.

[24] The only similarity to the Ostensive Model of Relevance (Campbell, 2000) is the exponentially increasing relevance weight applied to document representations at subsequent temporal positions.

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 set of results was never generated since I want to evaluate the performance of the model between searcher-defined query iterations. In the simulation searchers were modelled using a number of different strategies: (i) assume they only view relevant/non-relevant information, i.e., follow relevance paths from only relevant or only non-relevant documents, (ii) assume they view all relevant or all non-relevant information, i.e., follow all relevance paths from top-ranked relevant documents or top-ranked non-ranked documents, (iii) exhibit differing degrees of 'wandering' behaviour, i.e., try to view relevant information but also viewing different amounts of non-relevant information.

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

In this section the evaluation methodology is introduced. The system, corpus and topics used are described in Section 8.3.1. In Section 8.3.2 the techniques used to extract the relevance paths are described and in Section 8.3.3 the different searcher modelling strategies that use the relevance paths are described. In Section 8.3.4 the relevant distributions and correlation coefficients used to evaluate how well the models learn relevance are presented. The procedure and a description of the study are given in Section 8.3.5 and 8.3.6 respectively.

8.3.1 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 (SJMN 1991) document collection taken from the TREC initiative (Voorhees and Harman, 2000). This collection comprises 90,257 documents, with a mean average 410.7 words per document (including document title), a mean average 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' (see Chapter Five, Section 5.2.5) the component requires documents that contain around ten sentences. [25]

TREC topics 101-150 were used and the query was taken from the short *title* field of the TREC topic description. For each query the top 30 documents are used to generate relevance paths for use in the simulation. The number and nature of relevance paths chosen for the simulation is dependent on the simulation strategy employed, i.e., how the simulated searchers interact and how relevance paths are selected. The simulation assumes that searchers look at a subset of relevant paths, all relevant paths or a mixture of relevant and non-relevant paths. Non-relevant paths are assumed taken from non-relevant documents.

The simulation retrieves the top 30 results for each of the 50 TREC topics used as queries in this study; these results can contain both relevant and non-relevant documents. In some scenarios the simulation requires paths from only non-relevant documents, only relevant documents or a mixture of both. However, for some topics, there are no relevant documents in the top 30 results, making the execution of the latter two scenarios problematic. Therefore, when the simulation uses paths from relevant documents, it uses only those queries that have relevant top-ranked documents (i.e., 43 of the 50 topics have relevant documents in the top 30). There are non-relevant documents in the top 30 for all topics, so the same problem does not arise. I now explain how paths are extracted from top-ranked results for each topic.

8.3.2 Relevance Paths

In the simulation paths are extracted only from relevant documents, only from non-relevant documents or from a mixture of relevant and non-relevant documents, depending on the simulation strategy. Each document has a set number of representations and number of possible relevance path routes between these representations. In Table 8.1 all routes for all path types are shown. The final 'document' step is not included in the simulation since it is not used by the implicit feedback models at the search interface.

[25] Documents with only four sentences may result in low quality summaries and sentences in context comprised of other summary sentences, not new sentences that may contain useful alternate terms.

Table 8.1
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

For example, for viewing all five representations (first row of Table 8.1) there are 4 × 1 × 1 × 4 × 1 = 16 possible paths. The final column shows the total for each possible route. There are 54 possible relevance paths for each document. If all top 30 documents are used there are 1,620 (54 × 30) possible relevance paths per search topic. In the next section more details are given on how search scenarios that use these paths are deployed in the simulation.

8.3.3 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 *pre-modelled* situations. In this section styles of interaction that represent each of these situation categories are described in more detail. Paths and documents are considered synonymous unless otherwise stated.

8.3.3.1 Extreme Situations

Styles of interaction in this category represent extreme situations where *only* relevant or non-relevant 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 non-relevant information. They determine the best or worst expected performance of the models, depending on the paths or documents chosen.

8.3.3.1.1 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 non-relevant top-ranked documents this approach allowed the best and worst performing paths (and sets of paths) for each topic, and across all topics, to be identified. This can be useful to establish the attributes of good and bad relevance paths.

8.3.3.1.2 Subset of Paths

Searchers would typically not view all retrieved information. This strategy randomly selects a subset of paths used in the 'All Paths' situation. Paths are traversed in sequence and the effect across paths is cumulative. That is, unlike the 'All Paths' situation, the term scores in the term selection models are not reset after each path.

8.3.3.2 Pre-modelled Situations

The implicit feedback frameworks described in Chapters Six and Seven assume searchers will try to interact with relevant information, but accept they will inevitably also view information that is non-relevant. Pre-modelled situations model circumstances where searchers may view relevant and non-relevant 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 irrelevant. 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. The empirical findings of the user experiment presented in Part IV of this thesis suggests that non-relevant relevance paths contained fewer steps than relevant paths. As will be shown later in this thesis, the Checkbox system in that experiment allowed subjects to assess the relevance of each document representation in a relevance path and explicitly communicate their decisions to the retrieval system. Paths with no relevance assessments were shorter than those with at least one assessment, suggesting that irrelevant paths should be shorter in the simulation. Inferences made from interaction logs can assist in the development of richer simulated search strategies that can better approximate the interaction of real searchers. In Section 8.3.3.3 I use these data to model the length of relevance paths.

It is possible to vary how relevant (*R*) and non-relevant (*N*) paths are distributed to test how the models perform in different circumstances. The distribution method described in this section use previously traversed paths to select future paths.

8.3.3.2.1 Related Paths

This method selects paths that are related to those previously followed. The first path to be visited is chosen at random from the list of available paths. This path can be relevant or non-relevant. Subsequent paths are randomised in such a way that for ten 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 onwards. 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 introduced in Chapter Seven. 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). 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. During a search session searchers would typically follow a series of *related* relevance paths in a rational way, viewing only the most useful or interesting. This strategy attempts to simulate this activity.

In {*R*, *N*, *R*, *N*, *N*, *R*, *R*, *N*, *R*, *N*} the path at position two is non-relevant. To select the actual path all candidate non-relevant paths are ranked based on the product of their quality and similarity to the path at position one. The highest ranked path is chosen as the next step and the process repeats until ten paths have been visited in the order described. Pre-modelled 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 information.

8.3.3.3 Path Length Distribution

The modelled situations use empirical evidence to decide that relevance paths taken from irrelevant documents were short, i.e., three steps or less. However, it is possible to further analyse these results and derive another strategy that creates a distribution of path lengths across relevant and non-relevant paths. Data gathered from interactive experimentation with the Checkbox system in Part IV of this thesis allowed the construction of path length distributions. This system allowed subjects to explicitly mark document representations as relevant. In that experiment, relevance paths considered as relevant if one or more of its constituent representations were marked as relevant by experimental subjects. Table 8.2 shows how path lengths are distributed across relevant (containing marked representations) and non-relevant (containing no marked representations) relevance paths.

Table 8.2
Path length distribution in relevant and non-relevant paths (values are percentages).

Steps	Path type	
	Relevant	Non-relevant
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

From these results it appears that searchers interacted differently with relevant and irrelevant information. More specifically, it demonstrates that the paths were longer if they contained relevant information. The values in Table 8.2 can be used in pre-modelled situations to control the number of paths of each length used in the simulation. For example, if there are ten relevant paths and 0% wandering i.e., $\{R, R, R, R, R, R, R, R, R, R\}$, then their would be one path of length one (14.18% of 10), one path of length two (9.53% of 10), two of length three (18.95% of 10), three of length four (25.11% of 10) and three of length five (32.23% of 10). The number of paths of each length is rounded to the nearest integer. These path length distributions may be used to simulate the general behaviour 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 length of paths from irrelevant documents.

In all strategies, model performance is measured based on how the modified queries they 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 I describe the use of relevant distributions and correlation coefficients to measure such learning.

8.3.4 Relevant Distributions and Correlation Coefficients

A good implicit feedback model should, given evidence from relevant documents, learn the distribution 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, computed in the same way as Equation 7.5. After each iteration the extent to which the term lists generated by the implicit feedback 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. I 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 I 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. I use both correlation coefficients to verify learning trends.

8.3.5 Evaluation Procedure

The simulation creates a set of relevance paths for all relevant and non-relevant 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*) depends on the simulation strategy employed. After each iteration, I monitor the effect on search effectiveness and how closely the terms chosen by the model correlate 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*:

i. use SMART to retrieve document set in response to query (i.e., topic title) using an *idf* weighting scheme and record the initial precision values.

ii. identify relevant or non-relevant documents in the top 30 retrieved documents, depending on the experimental run and store in set *s*.

iii. select Top-Ranking Sentences from all documents in *s* using the approach presented in Chapter Three.

iv. create and store all potential relevance paths for each document in *s* (up to a maximum of 54 per document).

v. choose relevance paths or documents as suggested by the simulation strategy, setting *m* to the number chosen. The Java [26] random number generator is used where appropriate in selecting random paths or documents.

vi. for *each* of the *m* relevance paths/documents:

   a. weight terms in path/document with chosen model and rank terms based on weights.
   b. monitor correlation between terms and topic's relevant distribution.
   c. choose top-ranked terms and use them to expand original query.
   d. use new query to retrieve new set of documents.
   e. compute new precision values.

To better 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 the next section the simulated study is described.

8.3.6 Simulated Study

A study of how well each term selection model learned relevance and generated queries that enhanced search effectiveness is now presented. The models are tested in extreme and pre-modelled situations and each requires a different evaluation approach. The strategies used either the 43 'useable' topics (only paths from relevant documents or a mixture of relevant and non-relevant documents) or all 50 topics (only paths from non-relevant 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 was a reasonable number of additional terms for use in simulated experiments. Query expansion was used to test the marginal effectiveness of the model i.e., 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.

8.3.6.1 Extreme Situations

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

[26] http://java.sun.com

8.3.6.1.1 All Paths

This strategy uses all paths from the top 30 relevant documents and all paths from the top 30 non-relevant documents. A run of the simulation comprised 54*n* relevance paths, where *n* is the number of relevant/non-relevant documents. The correlation coefficients and search effectiveness were measured after each iteration. The effect of term scoring across consecutive paths is not cumulative. That is, paths were treated in isolation. The evaluation investigated performance differences of paths generated (e.g., best path/worst path).

8.3.6.1.2 Subset of Paths

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

8.3.6.2 Pre-modelled Situations

Three pre-modelled methods were tested in this study. Unlike the extreme situations these methods do 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 I ran the simulation ten times for each implicit feedback 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 pre-modelled 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.

8.3.6.3 Experimental Scenarios

In this section I describe the eight simulated scenarios that test the implicit feedback models in different situations. Table 8.3 shows these scenarios and the variables changed in each scenario. If a variable varies as part of a scenario a dot (•) is shown in the corresponding cell.

Table 8.3
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		•			•	•	•

Scenarios 3, 4 and 5 are each divided into scenarios ‘a’ and ‘b’. In ‘a’ paths are selected randomly whereas in ‘b’ a path length distribution is used to select paths. In each scenario all six implicit feedback models introduced earlier in this chapter are used to generate new queries. The resultant precision values and correlation coefficients are used to assess the performance of the models. In the next section I describe the results of the simulated study for each experimental scenario with each implicit feedback model.

8.4 Results

The study was conducted to evaluate a variety of implicit feedback models using searcher simulations. In this section I present results of the study for each simulation strategy. In particular I focus on results concerning search effectiveness and relevance learning. I 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. All uses of the term ‘average’ in the remainder of this chapter refer to the *mean average*.

8.4.1 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 54*n* paths, where *n* is the total number of relevant documents in the top-30 retrieved. In total, there were 15,174 paths (i.e., 54 × 281 [27]) across the 43 topics 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

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

the term selection model learned relevance. This allowed me to predict the ten best and worst performing paths and analyse why some paths were good and some were bad. In Tables 8.4 and 8.5 I show the average best and worst path performance for each of the six term selection models. Also included are 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 these tables I also show total number of terms in a path and in brackets the percentage of those terms that are stopwords (i.e., common words such as 'a', 'the' and 'of').

Table 8.4
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 8.5
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

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 maximise the rate of learning when shown relevant information, but minimise the negative effects when shown irrelevant information.

From these tests path length, the number of terms and percentage of those terms that were stop words have little influence over path performance. However the indicativity, or *quality*, appears different between good and bad performing paths. I 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 performs best, as they have the highest potential marginal gains in correlation coefficients and the lowest potential marginal losses for selecting random path from the set of all paths.

8.4.2 Scenario 2: All Non-Relevant Paths

This scenario was very similar to Scenario 1 but used paths from non-relevant documents rather than relevant. This was meant to model the situation where, by chance, searchers had viewed all paths from non-relevant documents. I use the top-ranked sentences from the non-relevant documents to create the representations that comprise the relevance path. I use these sentences as non-relevant information and not, say the bottom-ranked sentences from non-relevant 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 non-relevant.

In total there were 65,826 possible path routes (i.e., 54 × 1219 [28]) 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 chapter, the paths chosen from negative documents were assumed to be shorter than relevant paths. For each model, in Tables 8.6 and 8.7 I show the average path performance, the average number of terms and the proportion that are stopwords.

Table 8.6
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

[28] In total, there were 1219 non-relevant documents in the top 30 retrieved for all 50 search topics used.

Table 8.7
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 outperform the other term selection models. However, the wpq.doc model appears 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 are taken from non-relevant documents the indicativity of the relevant distribution (created from relevant documents) is lower than paths from relevant documents, shown in Tables 8.6 and 8.7. Also, for paths from non-relevant documents, there appears to be no association between path performance and relevant distribution indicativity.

In Scenario 2 (as in Scenario 1), the path length, the number of terms, number of those terms that were stopwords appears to have no effect on path performance. For Scenario 1 and Scenario 2 I did not measure precision after each path. Across relevant and non-relevant documents there were 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, I demonstrate a close relationship between the rate of learning and measures of precision; where it may not be practical to compute precision, correlation coefficients may be a reasonable approximation.

8.4.3 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.

8.4.3.1 Search Effectiveness

In Scenario 3a measured search effectiveness for each of the models through their effects on precision. Figure 8.1 shows the average *11-point precision* [29] values for each model across all iterations and 10 experimental runs. As the figure illustrates, precision increases as the

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

number of iterations increases. Figure 8.1 presents the actual precision values across all 20 iterations. The Jeffrey's Conditioning and Binary Voting Models outperform the other implicit feedback models, with large increases inside the first five iterations. Both models are 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 do not perform as well, but steadily increase until around 10 iterations where precision levels out.

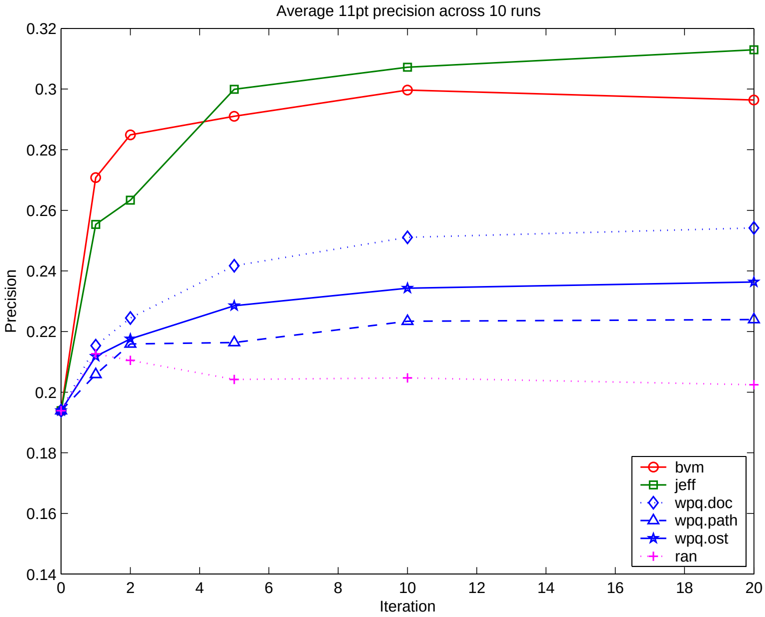


Figure 8.1. Average 11-point precision across 10 experimental runs in Scenario 3a.

Table 8.8 illustrates the marginal differences more clearly than Figure 8.1, showing the percentage change overall and the marginal percentage change at each iteration.

Table 8.8
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.

Model	Iterations									
	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	**38.0**	**+ 1.8**
wpq.doc	10.0	–	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	+ 0.9
ran	8.8	–	7.9	– 1.1	5.0	– 3.1	5.3	+ 0.2	4.2	– 1.1

As Table 8.8 shows, the largest increases in precision come from the Binary Voting Model and the Jeffrey's Conditioning Model. Although after 20 iterations the marginal effects of all models appear slight. The random model performs poorly, although still leads 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. My results show that choosing terms randomly from paths can slightly improve short queries.

The *wpq*-based models appeared to follow a similar trend. At each iteration a one-way repeated measures 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.

8.4.3.2 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; Figures 8.2 and 8.3 show the average Spearman and Kendall correlation coefficients across all 43 topics.

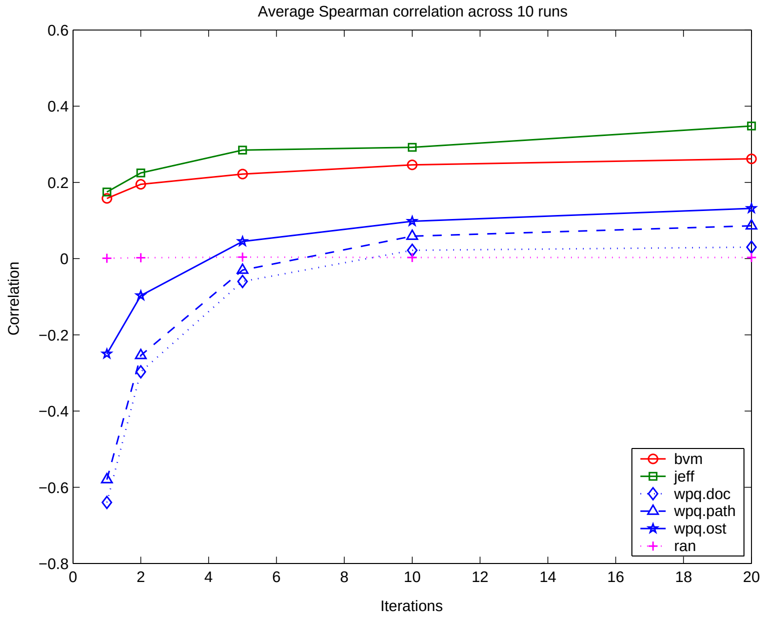


Figure 8.2. Average Spearman correlation coefficient across 10 runs in Scenario 3a.

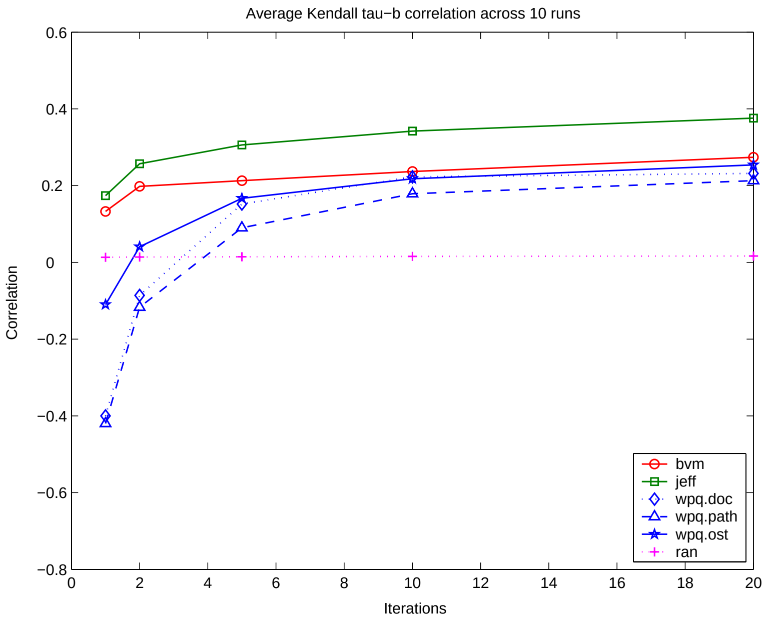


Figure 8.3. Average Kendall correlation coefficient across 10 runs in Scenario 3a.

Both coefficients follow similar trends for all implicit models. Again the Jeffrey's Conditioning and Binary Voting Model learn at a faster rate, with the Jeffrey's Conditioning Model performing best. The random model returns 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 tends to one, 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 obtain high levels of correlation after the first iteration, whereas the *wpq* models need more *training* to reach a level where the terms they recommend appear 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 results of findings of this scenario demonstrated little difference with the random paths approach used in Scenario 3a.

8.4.4 Scenarios 4a and 4b: Subset of Paths

Scenarios 4a and 4b, in a similar way to Scenarios 3a and 3b, use a subset of available paths. This scenario models the situation that may arise if, by chance, all information a searcher views is from documents that were non-relevant. It is reasonable to assume that searchers will view *some* information from non-relevant documents as they search. It is only in extreme situations where *all* the information they view is from non-relevant documents. These scenarios model such an extreme situation.

8.4.4.1 Search Effectiveness

I measured search effectiveness for each of the models through their effects on precision. Figure 8.4 shows the 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 outperform the other implicit feedback models. Although the increases in precision are small, the Jeffrey's Conditioning and Binary Voting Models seem better able to create effective search queries in situations where relevant information is difficult to find. That is, they seem better able to use paths from non-relevant documents to select terms for query modification. The other models do not perform as well, but steadily increase until around 10 iterations where precision levels out.

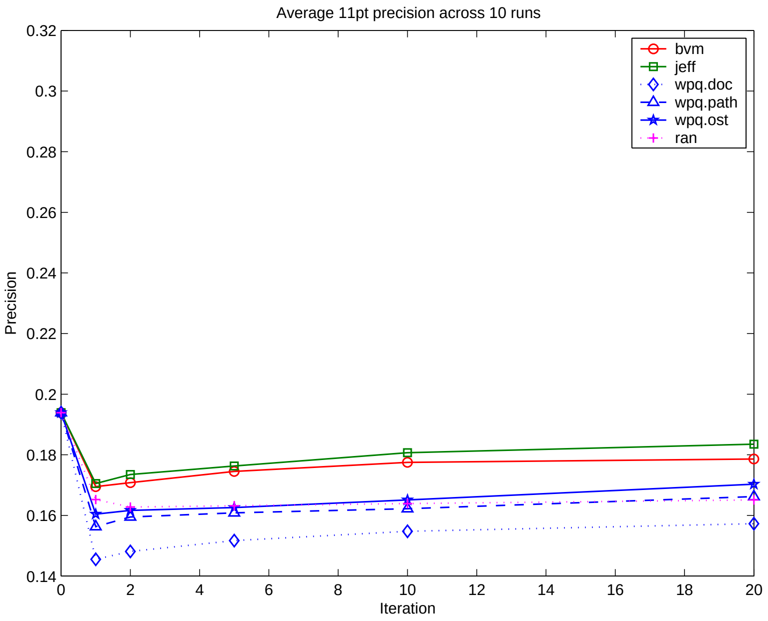


Figure 8.4. Average 11-point precision across 10 experimental runs in Scenario 4a.

The paths from non-relevant 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 Chapter Three. 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 titles in documents 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. Whilst all models appear to be affected by the presence of non-relevant information the Jeffrey's Conditioning and Binary Voting Models appear 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 increase precision slightly. With the exception of the *wpq.doc* model all models take terms from relevance paths that extract the most potentially useful parts of documents. Whilst the documents were classified by the TREC assessors as non-relevant 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 8.9 illustrates the marginal difference more clearly than Figure 8.5, showing the percentage change overall and the marginal percentage change at each iteration.

Table 8.9

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.

	Iterations									
Model	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

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 ≤ .05). As was demonstrated in Scenarios 3a and 3b, the Jeffrey's Conditioning and Binary Voting Models perform better than the other models in the first iteration. When presented with paths from non-relevant documents these models seem better able to extract useful terms. As is shown in Table 8.9, it is the first iteration that provides the overall increase in precision; after iteration one the marginal changes are similar for all models.

8.4.4.2 Relevance Learning

I 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; Figures 8.5 and 8.6 shows the average Spearman and Kendall correlation coefficients across all 50 topics.

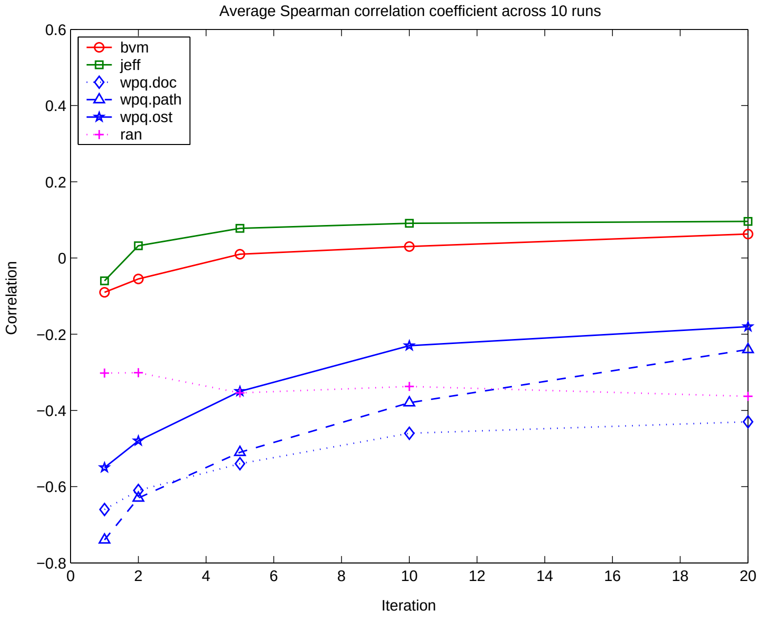


Figure 8.5. Average Spearman correlation coefficient across 10 runs in Scenario 4a.

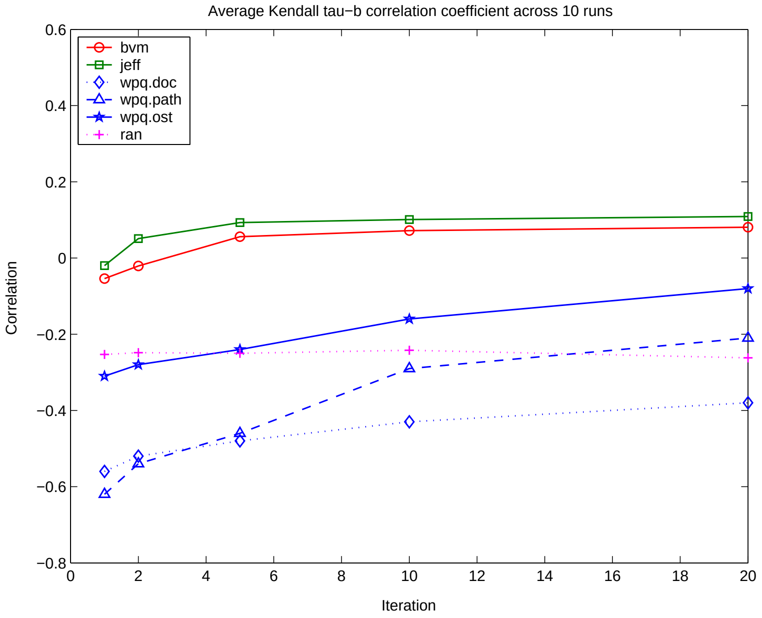


Figure 8.6. Average Kendall correlation coefficient across 10 runs in Scenario 4a.

The results show that in a similar way to Scenario 3, the models learn over time. However, since they are being shown information from non-relevant documents they do not learn the relevant distribution (composed of relevant documents) at as fast a rate and do not finish with as high a correlation as in Scenarios 3a and 3b. The random model returns a coefficient value close to zero with both coefficients in 3a and 3b. However, in this scenario it is lower, suggesting it starts at a low rate of learning and does not improve on this. The models based on *wpq* also perform poorly initially but improve gradually as the search proceeds.

In a similar way to 3b, Scenario 4b 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 threshold, 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.

8.4.5 Scenarios 5a and 5b: Related Paths

This scenario uses the 'Related Paths' approach described in Section 8.3.3.2.1 to select paths from relevant and non-relevant documents. Search effectiveness (monitored through precision) and relevance learning (measured through correlation coefficients) are monitored for different levels of wandering. In this section I summarise 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%). I present the actual values obtained for each of these levels in Appendix A. This approach is potentially more realistic than the experimental scenarios presented so far in this chapter, as it is conceivable that searchers will view irrelevant information as they search.

8.4.5.1 Search Effectiveness

As in previous scenarios the 11-point precision value was measured at iterations 1, 2, 5, 10 and 20. In Figure 8.7 I present the average precision value across all 10 runs and across all levels of wandering. The trend is the same as in earlier scenarios, with the Jeffrey's Conditioning and Binary Voting Models leading to overall increases in precision. However, because I introduce non-relevant 'noise' into the calculation, the overall increases in precision are not as large as in Scenarios 3a and 3b.

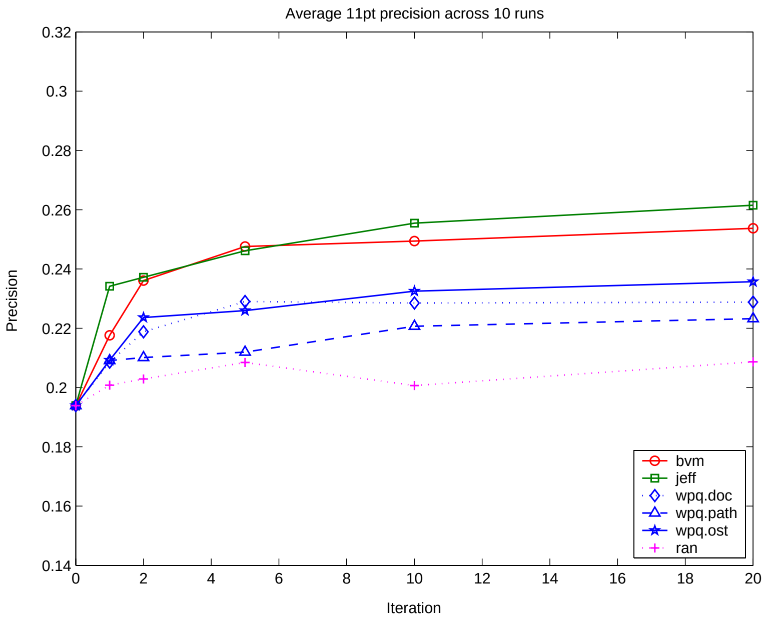


Figure 8.7. Average 11-point precision across 10 experimental runs in Scenario 5a.

The percentage change in overall and marginal precision for each of the models is shown in Table 8.10.

Table 8.10

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.

Model	Iterations									
	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 rises, increases in the level of precision drop. Viewing information from non-relevant documents (as Scenarios 4a and 4b demonstrate) is to reduce the overall effectiveness of all the term selection models. Nonetheless, the Jeffrey's Conditioning and Binary Voting Models still outperform the others.

8.4.5.2 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 Figures 8.8 and 8.9 respectively.

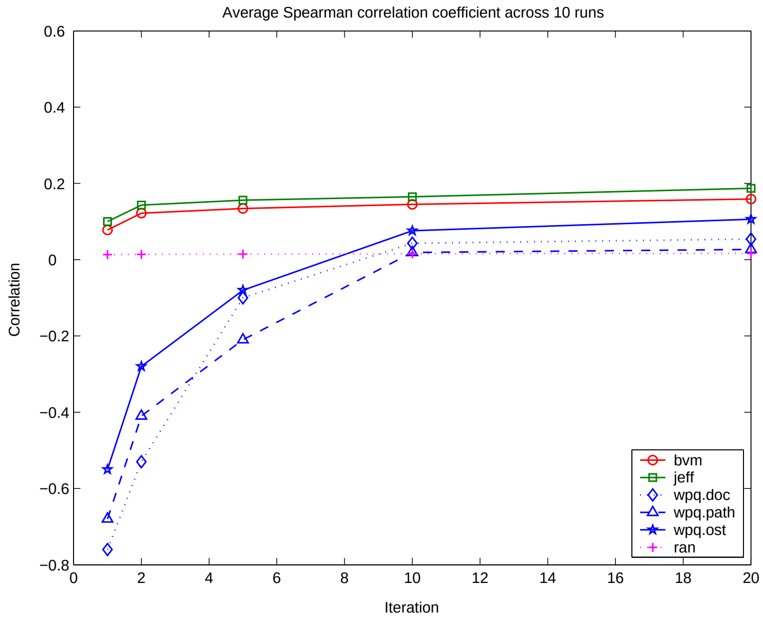


Figure 8.8. Average Spearman correlation coefficient across 10 runs in Scenario 5a.

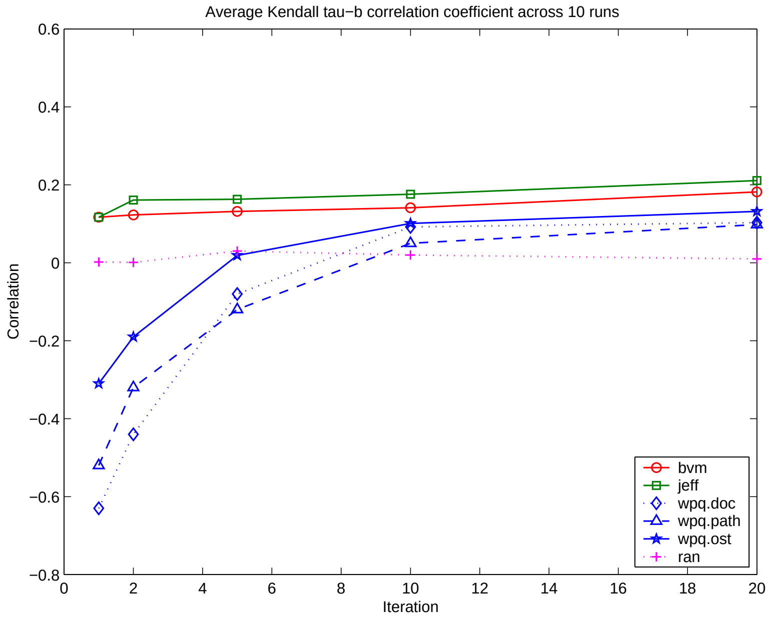


Figure 8.9. Average Kendall correlation coefficient across 10 runs in Scenario 5a.

Even though the models are shown potentially non-relevant information the results still demonstrate that the models are able to learn. However, their ability to do so is affected by the level of wandering. As wandering increases the rate at which the models learn relevance decreases. The actual correlation values for different levels of wandering are presented in Appendix B.

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 may not be as similar to the current path as other candidate paths, but had to be chosen to full the percentage quota of the distribution. The overall effectiveness of the models was reduced by around 5% by imposing the path length restriction. I present the actual values for Scenario 5b in Appendix C. In the next section I discuss this study's results.

8.5 Discussion

The implicit feedback models evaluated in this paper *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 does not perform as well as the Binary Voting Model at the start of the search. However, after five paths it creates more effective queries and from then on performs increasingly better than it. The Jeffrey's Conditioning Model uses prior evidence that is independent of the searcher's interaction. Initial decisions are made based on this prior evidence, and for the first few iterations it is reasonable to assume that this evidence still plays a part in term selection. However, as more evidence is gathered from searcher interaction the terms selected by the Jeffrey's Conditioning Model improve.

An advantage of the Binary Voting Model, and perhaps why it performs well in the initial stages is that it does not rely on any prior evidence, selecting terms based only on the representations viewed by the searcher. However, the lists of potential terms offered stagnates after 10 paths, since in the Binary Voting Model the effect of the scoring is cumulative, the high-scoring, high-occurrence terms, obtain a higher score after only a few initial paths and cannot be succeeded by lower-ranked terms in later paths. This often means that the same query is 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 me to test how well each model trained itself to recognise relevance. From the six models tested, the 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 in which the models learn relevance appears to match the rate in which they are able to improve search effectiveness.

The findings of the study show that the Jeffrey's Conditioning and Binary Voting Models are able to perform more effectively than the baselines when all the paths presented to them are from non-relevant documents (Scenarios 4a and 4b) and only a proportion of the paths are (Scenarios 5a and 5b). Whilst 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 non-relevant information is also shown. This is important in implicit feedback models as they assume a degree of relevance in all the information searchers view.

From the three models that implement different versions of the *wpq* algorithm, the wpq.doc model performed best for all relevant documents (Scenarios 3a and 3b) and worst for all non-relevant documents (Scenarios 4a and 4b). This model is 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 I have also shown that paths that lead to 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 iterations on all models, the marginal increases in precision and correlation reduce as more relevant information is presented. The models appear to reach a point of saturation at around 10 paths, where the benefits of showing 10 more paths (i.e., going to iteration 20) are only very slight and are 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.

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. In the next section I summarise this chapter.

8.6 Chapter Summary

In this chapter a simulation-based evaluation methodology called SIMPLE was presented and used to evaluate a variety of implicit feedback models. The models under test were ostensive in nature and use the exploration of the information space and the viewing of information as an indication of relevance. Six 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 granular representations of documents. The simulation passes the information it viewed to the implicit feedback models, which use this evidence to select terms to best describe this information. I 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 and the highest rate of learning.

Simulation experiments are a reasonable way to test the worth of implicit feedback models such as those presented in this chapter. However, whilst the simulation allowed me to benchmark model performance, evaluation with simulations is only formative and there is a need for further investigation of the best performing model when it is employed by real searchers engaged in IIR. In Part IV a user experiment is conducted of feedback systems that use the Jeffrey's Conditioning Model for term selection. In this chapter I have assessed the performance of the model objectively, using measures of search effectiveness and relevance learning. In the experiment in subsequent chapters, the performance of the model is assessed using human subjects.

Part IV

User Experiment

In Part III I described two implicit feedback frameworks: one heuristic-based and one probabilistic. Both approaches used searcher interaction with document representations to generate new query statements and estimate changes in the information needs of searchers. The part concluded with a simulation-based evaluation of different candidate implicit feedback models, including parts of the heuristic-based and probabilistic frameworks from earlier chapters. The probabilistic model based partly on Jeffrey's rule of conditioning performed best and was therefore selected as part of the experiment now presented. The experiment tests the value of the framework in detecting current information needs and tracking them over a search session, and the effectiveness of different types of interface support to communicate its decisions. Unlike the tests carried out in Part III, this experiment involves human subjects, and in addition to testing the probabilistic framework this experiment evaluates how much control searchers really want over in their interaction with the implicit feedback framework though the provision of relevance information, query reformulation and making search decisions.

Chapter 9

Experimental Methodology

9.1 Introduction

The simulation-based study in the previous chapter tested how well implicit feedback models improved search effectiveness and 'learned' what information was relevant. The study found that the term selection model based on Jeffrey's rule of conditioning outperformed the other models tested in a variety of information seeking contexts. In this chapter the value of the probabilistic framework described in Chapter Seven (of which the Jeffrey's Conditioning Model is part) is tested with human subjects. The framework includes components to estimate information needs and track changes in them over a single search session. The experiment also evaluates different forms of interface support for presenting the decisions the framework makes. Three search interfaces are evaluated that vary the amount of control searchers have over creating queries, providing relevance indications and making search decisions. In this chapter I describe the methodology used to evaluate the probabilistic framework and interface support mechanisms in all experimental systems. The chapter begins by describing two pilot studies, and then further describes the experimental methodology.

9.2 Pilot Testing

Two pilot tests were carried out prior to this experiment: one tested the a prototype content-rich interface and the heuristic-based framework described in Chapter Six, the second debugged the questionnaires and search tasks used the experiment described in this chapter. In the remainder of this section I describe each of these tests.

9.2.1 Pilot Test 1: Interface and Heuristic-based Framework

The first pilot test evaluated a prototype system developed based on the content-driven principles described in Part II. This tested the interface support mechanisms and the effectiveness of the heuristic-based implicit feedback framework described in Chapter Six. Two experimental interfaces were created and 24 experimental subjects were recruited. This test allowed me to evaluate a prototype version of the interface used in the experiment described later in this chapter. As a result, I resolved interface design issues, obtained a better understanding of subject interaction with such interfaces, and established the effectiveness of the heuristic-based implicit feedback framework. This test is described in more detail in Appendix D.

9.2.2 Pilot Test 2: Questionnaires and Search Tasks

This second pilot study debugged the questionnaires and the search tasks used in this experiment. Minor changes to the wording of questions in the questionnaires were made as a result of subject feedback. However, the main aim of this pilot test was to investigate the suitability and complexity of the search topics. In the main experiment subjects are required to choose three search tasks, one of high complexity, one of moderate complexity and one of low complexity. Subjects were presented with three task sheets, each containing six tasks on six topics. Subjects chose a task from each sheet, but could not choose the same topic more than once.

Borlund (2000b) suggested the most important factor in a good simulated situation was the degree to which the topic engaged the subject’s interest. Allowing subjects to choose tasks gave them more control over the search situation they were engaged in than simply allocating tasks to them on an arbitrary basis. In Pilot Test 1 I found that the level of interest in the search topic was the most important factor for experimental subjects when choosing one task over other alternatives.

Prior to starting the experiment, the task sheets were given to six randomly chosen volunteers. The volunteers were asked to read each of the tasks, place themselves in the simulated search scenario, and comment on the clarity and complexity of the task. These comments were informal and are not reported in this thesis. However, they did motivate slight changes in the wording of some tasks. In general, feedback on task complexity matched the categorisation used when developing the tasks. This was tested further in the main experiment and results are reported in later chapters.

In this section I have described two pilot tests that evaluate a prototype of the systems used in this experiment and debugged the questionnaires, search tasks and experimental procedures. In the remainder of this chapter I describe the methodology for the main experiment, beginning in the next section with the experimental systems.

9.3 Experimental Systems

Three experimental systems were developed to test these hypotheses. These systems varied in three ways: *relevance indication*, *query formulation* and *retrieval strategy selection* and used variations of interface components tested already in this thesis. A 'Checkbox' system (S_{Check}) allowed searchers to mark relevant items and use the items marked to create new queries. A 'Recommendation' system (S_{Recomm}) suggested additional query terms and retrieval strategies based on implicit relevance indications gathered from searcher interaction. An 'Automatic' system (S_{Auto}) automatically creates a new query and chooses the most appropriate retrieval strategies. No system gave subjects complete control over the terms used and search decisions taken. That is, all systems offered assistance in creating new queries, choosing how to use these queries, or both activities. Previous studies in IR have demonstrated that systems that offer feedback outperform systems where searchers are solely responsible for interaction decisions (Koenemann and Belkin, 1996; Beaulieu, 1997). I therefore felt it was unnecessary to include such a system did not offer any support in this experiment. These systems are described in more detail in Chapter Ten.

9.4 Equipment

I controlled the experiment from a laptop computer. The experimental systems ran on this computer and I sat next to computer for the duration of the experiment. An additional 21 inch monitor, a standard QWERTY keyboard and two-button optical mouse were connected to the laptop. [30] The experimental subject used these standard devices rather than those on the laptop, as shown in Figure 9.1. I felt these devices were more familiar to subjects than those on the laptop, which had a smaller display, a smaller keyboard and a touchpad for controlling the mouse pointer.

[30] The laptop computer had an AMD Athlon 2.4 GHz processor with 512 MB of RAM. The operating system was Microsoft Windows XP Professional and the Web browser used was Internet Explorer 6.0. All applications were written in Java, Dynamic HTML and JavaScript.

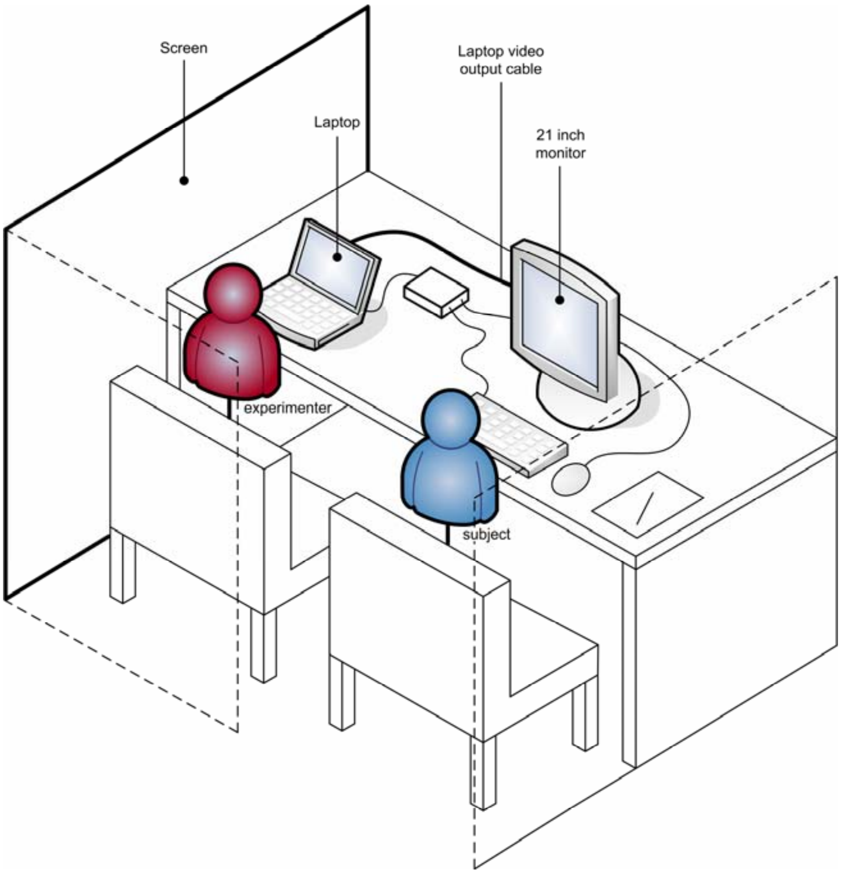


Figure 9.1. Equipment setup for the experiment.

Screens were positioned on three sides of the experimental location to block off noise and other distractions. I used the laptop to control the setup of experimental systems, control the construction of interaction log headers (described in Section 9.11) and observe subject interaction in an unobtrusive way. This also allowed me to intervene should there be any problems with the experimental systems. This intervention was limited only to occasions where technical problems prevented the subject from continuing with their search; I offered no other support.

9.5 Document Domain

The World Wide Web was used as the document domain for this experiment since subjects had experience interacting with Web documents, effective search systems were readily available and realistic search scenarios could be easily created. No restrictions were placed on the type of document that could be viewed or how far away from the experimental systems' result interface the subjects could browse. Restrictions were placed on whether external search systems (e.g., Google) could be used. These were seen as replacements for the experimental systems and were not permitted. Subjects were allowed to search within a document using the 'Find' function of the Internet Explorer browser. Many subjects used this function to locate keywords within a Web document.

9.6 Subjects

The experimental subjects were mainly staff and undergraduate and postgraduate students at the University of Glasgow. 48 subjects were recruited. Half were male and half were female. Subjects were paid £12 (approximately €18) for participating. In this section I describe how volunteers were recruited and how the final set of subjects was selected.

9.6.1 Recruitment

Recruitment was targeted at two groups of subjects; *inexperienced* and *experienced*. In a related study, Holscher and Strube (2000) showed that experienced and novice Web searchers conduct their searches differently. Since the Web has a heterogeneous user population it is important to investigate how well the techniques I propose perform for different subject groups. I define the subject groups as:

i. **Inexperienced:** infrequent computer users, inexperienced searchers.
ii. **Experienced:** frequent/professional computer users, experienced searchers.

Subjects were not classified into their groups until after they had completed an 'Entry' questionnaire that asked them about their search experience and computer use. Subjects were recruited using electronic mails and advertisements per the ethics code of the Faculty of Information and Mathematical Sciences, University of Glasgow. These recruitment methods yielded of a pool of 156 interested volunteers. In the next section I describe how 48 subjects were chosen from this pool.

9.6.2 Selection

The name and email addresses of each subject were stored electronically. The list of subjects was divided based on volunteer gender (male 63.38%, female 36.62%). Subjects were sampled at random from these groups until 24 males and 24 females were chosen and notified through electronic mail. They were asked to visit a Web page containing an experimental timetable, select a small set of the most convenient times and respond via email. Experimental time slots were allocated based on subject preference and availability of suitable times. A time slot was allocated and a confirmation email sent.

Experimental subjects were assigned a unique experiment identifier in the range 101-148. This identifier was used during experimental data capture and analysis.

9.6.3 Subject Demographics and Search Experience

The average age of the subjects was 22.83 years (maximum 51, minimum 18, standard deviation = 5.23 years). Three quarters had a university diploma or a higher degree and 47.91% of subjects (23) had, or were pursuing, a qualification in a discipline related to Computing Science. The subjects were a mixture of students, researchers, academic staff and others. They had different levels of computing and search experience.

The subjects were divided into two groups – *inexperienced* and *experienced* – depending on their computing and search experience, how often they searched and the types of searches they performed. All were familiar with Web searching, and some with searching in other domains. The division of these groups was potentially problematic as subjects may not give an accurate account of their experience level. Table 9.1 shows the composition of each group and the differences between groups.

Table 9.1
Inexperienced and Experienced subject characteristics.

Factor	Inexperienced	Experienced
Number of subjects	24 (12 male, 12 female)	24 (12 male, 12 female)
Average search frequency	‘Once or twice a week’	‘Many times a day’
Use point-and-click interfaces	‘Frequently’ (3.58)	‘A lot’ (4.96)
Use Web search engines	‘Frequently’ (4.08)	‘A lot’ (4.92)

Subjects were asked to complete Likert scales asking how much experience they had with point-and-click interfaces, such as Microsoft Windows, and Web search engines. These results are reported in the last two rows of Table 9.1. The Likert scale values are in the range 1 to 5, where a higher value corresponds to more experience. The differences between subject groups were significant with a Mann-Whitney Test. [31]

Subjects were also asked to indicate which Web search engines they used and complete semantic differentials on how ‘easy’/‘difficult’, ‘stressful’/‘relaxing’, ‘simple’/‘complex’ and ‘satisfying’/‘frustrating’ the general use of these search engines was. This was potentially a good indicator of experience levels as I would expect subjects with more experience to be

[31] Experience with point-and-click interfaces, $U(24) = 441$, $p < .001$, experience with Web search engines, $U(24) = 396$, $p = .013$.

more competent searchers. Table 9.2 showed the average differential responses and the significance of the differences between subject groups with a Mann-Whitney Test.

Table 9.2
Search engine use (scale from 1 to 5, lower = better).

Differential	Inexperienced	Experienced	Significance$^{\alpha}$
easy	2.29	1.50	.004
relaxing	2.63	2.46	.475
simple	2.13	1.63	.045
satisfying	2.46	2.46	.156

$^{\alpha}$ with a Mann-Whitney Test, *U*(24).

The results show that those subjects classified as ‘experienced’ found using Web search engines significantly easier than the inexperienced group; to a certain extent this validated the subject classification. In the next section I describe the search tasks given to experimental subjects.

9.7 Tasks

In this section I discuss the search tasks attempted by experimental subjects. Tasks were divided into three categories and within these categories into six search topics. The tasks were designed to encourage naturalistic search behaviour by experimental subjects. I wanted subjects to interact with the experimental systems as though they were performing their own search. To do this, the tasks were placed within simulated situations as proposed in Borlund (Borlund, 2000b; 2000a). The technique asserts that searchers should be given search scenarios that reflect and promote a real information seeking situation. Figure 9.2 shows an example simulated situation.

Simulated Situation

Simulated work task situation: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Indicative request: Find for instance something about future employment trends in industry, i.e., areas of growth and decline

Figure 9.2. Simulated situation taken from Borlund (2000a).

Simulated situations can be composed of two parts: the simulated work task situation and an indicative request. The simulated work task situation is a short 'cover-story' designed to provide context for a search. The indicative request is an indication, rather than an instruction, of how a search may be initiated. Previous studies have shown that the indicative request is not required for the simulated situation to engage the subject in the search and to promote natural searching behaviour on the part of the subject (Borlund, 2000a).

The simulated situations, such as that shown in Figure 9.2, are intended to achieve two main objectives. First, they promote a simulated information need in a subject. That is, the simulated situation should engage the subjects in the search by the identification of the searcher within the situation. As in Pilot Test 1, I offer subjects a choice of search tasks to go some way to ensuring they choose tasks of interest to them and can identify with the topic of the search. In Pilot Test 2, these tasks were tested for differences in their difficulty; no differences were found.

Second, the simulated situations position the search within a realistic context. The situation allows the experimental subject to provide his or her own interpretation of what information is required and allows them to develop the information need naturally. They permit a dynamic interpretation of relevance by experimental subjects. In forthcoming sections I describe the task categorisation and the search topics.

9.7.1 Task Categories

The tasks in this experiment were divided into three categories. Tasks were categorised based on their complexity and tried to encourage different types of information seeking behaviour. The aim of this approach was to create different types of needs to see how well the experimental systems performed for these differing types and to hopefully elicit different subject behaviours. The six stage Information Search Process (ISP) model (Kuhlthau, 1991) forms the basis of the task selection. I do not choose six task categories that correspond with the six stages in the ISP, but instead to the three types of searcher interaction that the model predicts; *background seeking*, *relevant seeking* and *relevant and focused seeking*. Through varying their complexity, this categorisation at least aims to encourage the types of interaction I would expect to see at each stage, in the hope that it may give a handle on what aspects of the search process each experimental system supports well, and what parts they do not. In earlier work (White *et al.*, 2003b) I proposed four categories of Web search; *fact search*, *decision search*, *search for a number of items* and *background search*. In an earlier study, Byström and Järvelin (1995) describe five task categories based on their complexity and *a* *priori* determinability. The *a priori* determinability measures how well the searcher can determine the required task inputs (information necessary for their search), processes (how to find the required information) and outcomes (how to recognise the required information) based on the initial task statement. Through increasing the uncertainty associated with each of these factors an experimenter can control the complexity of the task. Table 9.3 shows the relationship between the ISP categorisation used in the experiment and this related work.

Table 9.3
Task categorisation and related work.

Related Work	Task category		
	Pre-focus	Focus formation	Post-focus
Information seeking behaviour (Kuhlthau, 1991)	*background*	*relevant*	*relevant or focused*
Task type (White *et al.*, 2003b)	*background*	*decision*	*fact* and *search for a number of items*
Task complexity (Byström and Järvelin, 1995)	*known, genuine decision task* and *genuine decision task*	*normal decision task*	*normal information processing task* and *automatic information processing task*

To create the *pre-focus*, *focus formation* and *post-focus* task categories I varied the number of potential information sources and type of information required to complete a task (Bell and Ruthven, 2004). Six search topics were chosen for the experiment and a pre-focus, focus formation and post-focus version of each category was created. In the next section I describe these topics.

9.7.2 Search Topics

Six search topics were tested in Pilot Test 2 and used in this experiment. The topics were chosen to be of general interest to participants and reflect searches they may be likely to perform. The simulated work task situations used in this experiment were tailored towards the information environment and the group of test persons. Borlund (2003) recommends that this tailoring is to include:

i. A situation which the test persons can relate to and in which they can identify themselves;
ii. A situation that the test persons find topically interesting, and;
iii. A situation that provides enough imaginative context in order for the test persons to be able to relate and apply the situation.

Tailoring of simulated work task situations is important in order to gain a trustworthy behaviour and IR interaction from experimental subjects. Table 9.4 shows the topic titles for the six search topics used.

Table 9.4
Titles of search topics used during experiment.

1. Applying to university	4. Third generation phones
2. Allergies in the workplace	5. Internet music piracy
3. Art galleries in Rome	6. Petrol prices

For each of these topics three search tasks were created to match the *pre-focus*, *focus formation* and *post-focus* task categorisation. Subjects chose one pre-focus, one focus formation, and one-post focus task. They choose tasks from a different search topic each time and were not allowed to choose more than one task for a particular topic. This minimised task learning effects. The search tasks are included in Appendix F.3, where Task A is the high-complexity ‘pre-focus’ task, Task B is the moderate complexity ‘focus formation’ task and Task C is the low complexity ‘post-focus’ task. In the next section I describe how tasks were allocated to subjects.

9.7.3 Task Allocation

Borlund (2000a) conducted a feasibility test and revealed a ‘significant pattern of behaviour’ amongst experimental subjects in the way they carried out the relevance assessments of the retrieved documents when using simulated work task situations. For this reason an experimental design was used that could reduce the likelihood that the use of one system or attempting one task, influenced the next task-system variation. A Graeco-Latin square design was used (Tague-Sutcliffe, 1992), that rotated both experimental systems and tasks.

Table 9.5 shows the experimental design. The factors in the table are the *tasks categories* ($T_{A\text{-}C}$) and the *experimental systems* (S_{Check}, S_{Recomm}, S_{Auto}).

Table 9.5
Graeco-Latin square experimental block design.

Subject	System/Task order		
	1	2	3
1	S_{Check} , T_A	S_{Recomm} , T_B	S_{Auto} , T_C
2	S_{Auto} , T_B	S_{Check} , T_C	S_{Recomm} , T_A
3	S_{Recomm} , T_C	S_{Auto} , T_A	S_{Check} , T_B

This square represents a block of subjects. There are 16 similar blocks of three subjects in the experiment (i.e., 16 × 3 = 48). In the next section I describe the experimental procedure.

9.8 Procedure

Each subject was asked to attempt each of the search tasks they had chosen. The order in which topics were presented, and the choice of which system a subject used for each search, was determined by the randomised experimental matrix given in the previous section. Experiments lasted between one-and-a-half and two hours, dependent on the amount of time required to complete questionnaires. Subjects were provided with light refreshments and were offered a five minute break after the first hour.

For each experiment the following steps were followed:

i. Subjects were welcomed and asked to read the introduction to the experiment provided on an 'Information Sheet' (Appendix F.1). This set of instructions was developed to ensure that each subject received precisely the same information. Subjects could retain the information sheet after the experiment.

ii. Subjects were then asked to sign two copies of a consent form, one for my attention, and one on the reverse of the 'Information Sheet', for the subject to keep.

iii. Subjects were then asked to complete an 'Entry' questionnaire (Appendix F.2). This elicited background information on the subject's education, previous general search experience, computer use experience and Web search experience.

iv. Subjects were given a tutorial on all experimental systems, followed by a training topic. The training topic was the same for all subjects and is included in Appendix F.3. This training topic gave subjects a chance to familiarise themselves with the interface components of the experimental systems. More details on subject training are given in Section 9.9.

v. Once comfortable with the training system subjects were given the first task sheet and asked to select one search task from the six in the allotted task category. No guidelines were given to subjects about the criteria to use when choosing a task.

vi. After selecting the task, subjects were asked to perform the search it required. They were given 15 minutes to search and could stop early if they were unable to find any more relevant information.

vii. After completing the search (either successfully or otherwise), the subject was asked to complete the 'Search' questionnaire (Appendix F.2).

viii. The remaining task sheets were given to subject, following steps v. – vii. Since the search topics were the same on all three task sheets subjects were not allowed to choose the same topic as attempted in a previous search. Subjects were offered a five minute break after the first task (around halfway through the experiment).

ix. At the end of the experiment, the subject was asked to complete the post-experiment 'Exit' questionnaire (Appendix F.2) and an informal post-experiment interview was conducted.

The 'Search' and 'Exit' questionnaires were designed based on the research questions that motivated the experiment, described in Section 9.12. In the next section I provide more details on how experimental subjects were trained.

9.9 Training

Since the experimental systems were unfamiliar to subjects, they received pre-search training on how to use them. A short time, around 30 minutes was allocated for training at the start of the experiment. The training session was broken down into a series of stages:

i. I explained the purpose of the systems i.e., that they all tried to improve the quality of the subject's query and some tried to select new search decisions on the subject's behalf.

ii. Subjects were introduced to the search interface components that appeared in all systems (e.g., top-ranking titles, pop-up summaries). I used printed screenshots of each of the three experimental systems to help describe these interface components.

iii. I gave subjects a live demonstration of each system using the same search query, 'information'.

iv. A training task (Appendix F.3) was issued and subjects were given the chance to attempt this task on a training system with no feedback (similar to Koenemann and Belkin (1996)). The training task gave subjects an opportunity to use the system in a realistic information seeking context and become accustomed to the interface features.

v. The training session stopped once subjects felt comfortable using the systems.

Subjects were allowed to comment or ask questions at any point during the session. Due to the large number of experimental participants and the relatively short duration of the experiment, 30 minutes was the maximum time afforded to each subject. In all cases this appeared sufficient for subjects to familiarise themselves with the systems.

9.10 Questionnaires

Questionnaires were the main method used to elicit subject opinion during the experiment. The questionnaires were typically divided up into a series of sections that contained questions on the same aspect of the search (e.g., 'Search Process', 'Interface Support'). To help the subject complete the questions, some introductory text was given at the start of each section. Figure 9.3 gives an example of such text from the 'Search' questionnaire.

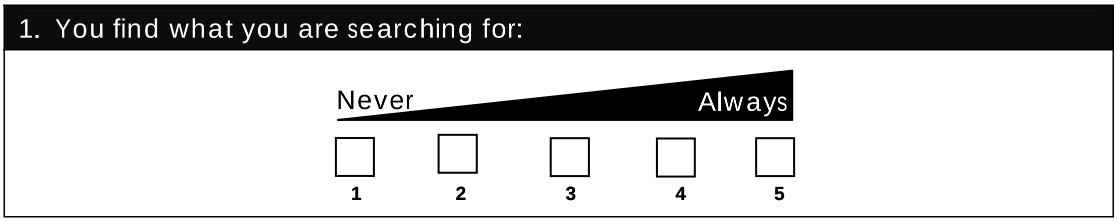
Relevance Assessment
The Automatic and Interactive systems assumed that much of the information you viewed was relevant. In the Checkbox system you explicitly marked relevant items.

Figure 9.3. Example introductory sentence (taken from 'Search' questionnaire).

Three questionnaires were developed and distributed to experimental subjects at various points in the search: 'Entry', 'Search' and 'Exit'. These questionnaires are included in Appendix F.2 and contained three styles of question; *Likert scales*, *semantic differentials* and *open-ended questions*. In this section each style is explained and examples provided.

9.10.1 Likert Scales

The Likert scaling technique presents a set of attitude statements. Subjects are asked to express agreement or disagreement on a five-point scale. [32] Each degree of agreement is given a numerical value from one to five. A total numerical value can be calculated from all the responses received. Figure 9.4 shows an example Likert scale taken from the 'Entry' questionnaire.



1. You find what you are searching for:

Never Always

□ □ □ □ □
1 2 3 4 5

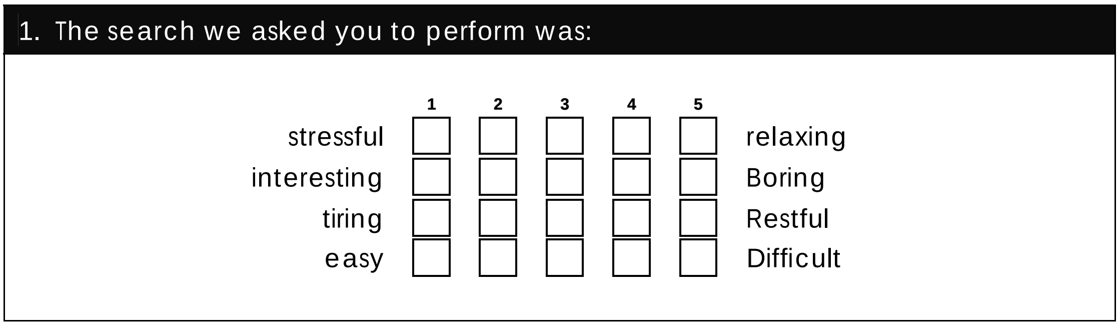
Figure 9.4. Example Likert scale (taken from 'Entry' questionnaire).

Likert scales are designed to show a differentiation among respondents who have a variety of opinions about an *attitude object* (i.e., anything that the subject may find good or bad), in this case how often they find what they are searching for.

[32] A five-point scale was preferred to seven or nine point scales as it made the analysis of subject opinion simpler and allowed trends in the results to be more easily identified.

9.10.2 Semantic Differentials

Another type of structured question is one that provides pairs of antonyms and synonyms, together with five-step rating scales. The word pairs refer to an attitude object, and respondents are asked to check one of the positions on each continuum between the most positive and negative terms. This type of scale is called a *semantic differential*. Figure 9.5 exemplifies a set of four semantic differentials.



1. The search we asked you to perform was:

	1	2	3	4	5	
stressful	□	□	□	□	□	relaxing
interesting	□	□	□	□	□	Boring
tiring	□	□	□	□	□	Restful
easy	□	□	□	□	□	Difficult

Figure 9.5. Example set of semantic differentials (taken from 'Search' questionnaire).

In this example, as in all differentials in the experimental questionnaires, the positive and negative terms are reversed in consecutive attitude objects. This ensures that subject attention does not waver when completing the questionnaires.

9.10.3 Unstructured Questions

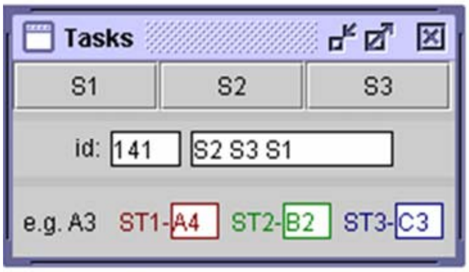
In unstructured questions subjects were given the chance to freely reply without having to select one of several provided responses; these questions can be described as 'open-ended'. They are useful for revealing reasons why subjects feel the way they do and giving them a chance to comment freely on aspects of the system, the task or the experiment in general.

Subjects were issued with an 'Information Sheet' at the start of the search that showed them completed examples of Likert scales and semantic differentials. It was assumed that subjects would not need instructions on answering unstructured questions.

During the experiment, system logging recorded search activity at the interfaces to the experimental systems. In the next section I describe the logging procedure used.

9.11 System Logging

Log files were named based on the subject's unique identifier, the system and task attempted. The log file contains a header, which is written before any interaction. This contained the subject identifier, the task being attempted, the experimental system being used and the date and time of the experiment. Prior to starting the each search task I created this header using a small Java application. The interface to this application is shown in Figure 9.6. It was not important that this interface was intelligible to experimental subjects as only I used it. The buttons S1, S2 and S3 can be used to clear system log files, the 'id' boxes contain the subject identifier and the order in which systems are used. In Figure 9.6, subject 141 is using S2 then S3 then S1. The search topic (ST) boxes contain the identifier of the search category/topic attempted (e.g., A4 is the fourth topic on the high complexity task sheet).



id: 141 S2 S3 S1

e.g. A3 ST1-A4 ST2-B2 ST3-C3

Figure 9.6. Java application for log header construction.

All searcher interaction with the experimental systems was also logged as a '<event> <timestamp>' pair and the timestamp was written as the number of milliseconds elapsed from midnight, January 1, 1970. This is a Java default and allowed times to be easily parsed and compared. Details of the tags used to denote the events and an excerpt from the log files are included in Appendix G.

The location of the mouse pointer is also logged every 0.25 seconds, and the locations of any mouse clicks are also recorded. From this log data I can analyse which parts of the interface subjects interact with and where they spend the most time. System usage data of this nature is useful for tracking exactly how subjects interact with these systems. In the next section I describe the experimental hypotheses tested during this experiment.

9.12 Hypotheses

The purpose of this experiment is to investigate the effectiveness of different forms of interface support for facilitating the use of relevance feedback in interactive search environments and the probabilistic framework described in Chapter Seven. The framework is used to modify queries and select retrieval strategies based on relevance feedback provided by the searcher. This feedback can be *implicit* (inferred by the system from interaction) or *explicit* (provided intentionally to the system by the searcher); different experimental systems offer different ways of indicating what information is relevant.

This experiment investigates which form of interface support searchers prefer, the ability of the probabilistic framework to choose worthwhile terms and the appropriateness of the new retrieval strategies chosen or recommended. In this section the experimental hypotheses are described. These are:

Interface support (Hypothesis 1)

Subjects like the interface support provided by the experimental systems and find that it facilitates effective information access.

Information need detection (Hypothesis 2)

Subjects find the terms chosen by the probabilistic implicit feedback framework valuable and worthwhile.

Information need tracking (Hypothesis 3)

Subjects find the retrieval strategies chosen by the probabilistic implicit feedback framework valuable and worthwhile.

The hypotheses are analysed in three ways. The first examines the subjects' overall search behaviour; this analysis looks for changes in how subjects searched on the experimental systems. The second examines the search effectiveness of the three systems; on which system did the subjects have a most effective search? Finally I shall examine the subjects' perceptions of the three systems; did the subjects prefer one system over the others?

9.13 Sub-hypotheses

It is possible to divide the experimental hypotheses provided in the previous section into a number of sub-hypotheses to make the capture and analysis of data more straightforward. In this section each set of sub-hypotheses are described.

9.13.1 Hypothesis 1: Interface Support

Five aspects of the interface support offered by the experimental systems were tested in this experiment:

Relevance Paths and Content (Hypothesis 1.1)

Subjects find the information presented at the interface useful.

Term selection (Hypothesis 1.2)

Subjects want control in formulating new queries.

Retrieval strategy selection (Hypothesis 1.3)

Subjects want control in making search decisions.

Relevance assessment (Hypothesis 1.4)

Subjects want the experimental system to infer relevance from their interaction.

Notification (Hypothesis 1.5)

Subjects find system notifications helpful and unobtrusive.

9.13.2 Hypothesis 2: Information Need Detection

This hypothesis assesses the effectiveness of the information need detection part of the probabilistic framework. To test it, subject opinion on the terms chosen by the term selection model was elicited. I divide the hypothesis into two sub-hypotheses based on their *value* (can be helpful during a search) and *worth* (is correct and accurate).

Value (Hypothesis 2.1)

Query modification terms chosen by the framework are relevant and useful.

Worth (Hypothesis 2.2)

Query modification terms chosen by the framework approximate subject information needs.

9.13.3 Hypothesis 3: Information Need Tracking

The information need tracking component of the system looked for changes in the information needs of searchers as they searched. The information need tracking component is tested via subject perceptions of the retrieval strategy selected by the system. That is, the component is evaluated through subject perceptions of the resultant *search strategy*, not the perceived extent of the change. There are two sub-hypotheses that, in a similar way to Hypothesis 2, are based on the value and worth of the component:

Value (Hypothesis 3.1)

The retrieval strategies chosen by the framework are beneficial.

Worth (Hypothesis 3.2)

The retrieval strategies chosen by the framework approximate changes in the information needs of subjects.

9.14 Chapter Summary

In this chapter the methodology has been presented for a user experiment to: (i) investigate interface support mechanisms to assist users of information retrieval systems and (ii) evaluate the effectiveness of the probabilistic implicit feedback framework in realistic search environments. The hypotheses for the experiment have been introduced and the document domain, tasks, subjects and experimental procedure have been described. In this chapter, the experimental systems used to test the hypotheses were briefly introduced. In the next chapter these systems are described in more detail.

Chapter 10

Experimental Systems

10.1 Introduction

Three experimental systems were created to test the hypotheses proposed in the previous chapter. The systems vary subject control over three main classes of decisions that users of such systems must make: selecting query terms, indicating relevance and making new search decisions. The experimental systems were: (i) a system that allowed subjects to directly communicate what information was relevant, provided support in creating new queries and allowed searchers to decide how these queries were used, (ii) a system that gathered relevance indications through implicit feedback, recommended new queries and made recommendations on how these queries should be used, and (iii) a system that used implicit feedback, automatically refined the query and made search decisions on query use on the subject's behalf. Each system offers different types of interface support, and where appropriate uses the techniques described in Chapter Seven. In this chapter I describe the experimental systems, their similarities and their differences.

10.2 Overview of Systems

The systems developed were interfaces to Web search engines that provided added support in creating search queries and making search decisions (i.e., re-searching the Web, reordering document lists and reordering lists of Top-Ranking Sentences). The names given to the systems during the experiments were based on their distinguishing features. The three experimental systems and *search activities* on each were:

i. **Checkbox:** searchers control relevance indication and query generation; searchers control query word selection; searchers control query execution.

ii. **Recommendation:** searchers delegate relevance indications and query generation; searchers control query word selection; searchers control query execution.

iii. **Automatic:** searchers delegate relevance indication and query generation; searchers delegate or control query word selection; searchers delegate or control query execution.

A summary of the responsibilities for all search activities is given in Table 10.1.

Table 10.1
System and subject responsibilities for search activities.

	System		
Search Activity	Checkbox	Recommendation	Automatic
Query Modification	System and Subject	System and Subject	System
Relevance Indication	Explicit	Implicit	Implicit
Retrieval Strategy Selection	Subject	System and Subject	System

The role of the subject in query modification is different in the Checkbox and Recommendation systems. In the Recommendation system they choose additional terms from those *recommended*; if a term is irrelevant subjects can ignore it. The Checkbox system selects additional terms and appends these to the original query in an editable text box. The subject is then responsible for retaining or removing terms to formulate the new query; if a term is irrelevant searchers have to delete it.

The experimental systems share a number of underlying features and differ in those necessary to test the research hypotheses outlined in the previous chapter. The aim of this thesis was not to develop an optimal search interface. The interfaces I constructed were developed for experimental purposes and were sufficient to allow an investigation of implicit feedback and interface support mechanisms. The probabilistic framework described in Chapter Seven is used by all systems to make decisions about query terms and, in the Recommendation and Automatic systems, to select retrieval strategies. In the next section the similarities and differences between the experimental systems are described.

10.3 Similarities and Differences

The systems share many features and differ in only a few. The differences between systems are limited to those necessary to test the research hypotheses.

10.3.1 Similarities

In this section the system features common to all three systems are described. Among other things, the systems share the same architecture for retrieving documents and selecting Top-

Ranking Sentences, general interface components, term selection model and method for scoring sentences and documents.

10.3.1.1 Retrieval Architecture

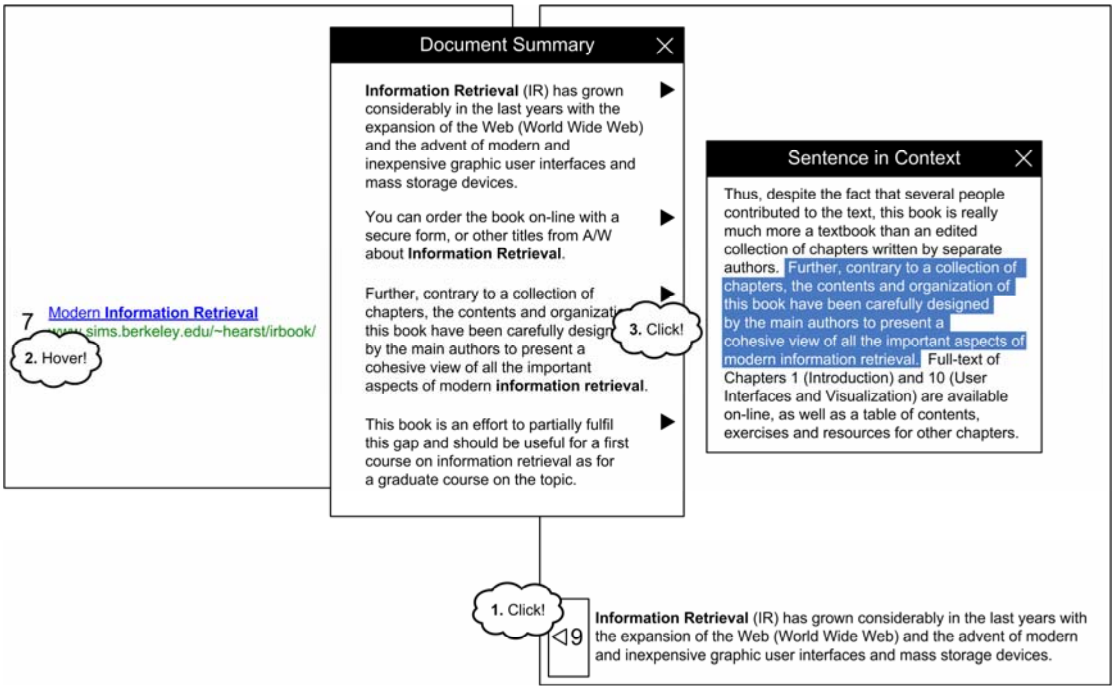
The same retrieval architecture underlies each of the three systems and is described in Chapter Three. All systems are implemented in Dynamic HTML (DHTML) and the client-side code for all systems is written in JavaScript. A submitted query is passed to the Google commercial Web search engine and the top-ranked documents are retrieved and the Top-Ranking Sentences selected. Google was chosen for the size of its index, the frequency with which this index is updated and the existence of a Java Application Programming Interface that allowed me to easily query the search engine. [33] The best sentences from all top-ranked documents are used to construct a list of Top-Ranking Sentences, presented to the searcher at the interface. A term space containing all unique terms in the most relevant documents is also constructed. [34] This space is used by the Jeffrey's Conditioning Model; each term in the space is considered a candidate for query modification.

10.3.1.2 Interface Components

The interfaces to the experimental systems in this experiment used titles, summaries and sentences as described in Chapter Five and in Pilot Test 1. However, unlike the interfaces used in Pilot Test 1 these interfaces use mouse clicks *on* search results rather than movements *over* search results as an indication of the relevance. Clicks show the subject the next step in the relevance path or open Web documents. Since the subject must act 'explicitly' (although not for the purpose of communicating relevance) each of these actions are assumed to be more reliable indicators of subject interests than mouse movements. A click represents a conscious effort by the subject and a break in their cognitive processes; clicks are normally intentional and can therefore be more reliable implicit relevance indicators than mouseovers. With mouseovers it can be difficult to determine what actions are intentional and which are accidental, arising through the movement of the mouse to another part of the screen. To follow a relevance path subjects must 'hover' over representations for a short period of time and click arrows next to representations as shown in Figure 10.1.

[33] http://www.google.com/apis/

[34] In the same way as Chapter Four, query-relevant Top-Ranking Sentences were selected from the top 30 retrieved documents to ensure the systems responded to the subject in a timely manner.



Document Summary

Information Retrieval (IR) has grown considerably in the last years with the expansion of the Web (World Wide Web) and the advent of modern and inexpensive graphic user interfaces and mass storage devices.

You can order the book on-line with a secure form, or other titles from A/W about **Information Retrieval**.

Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors to present a cohesive view of all the important aspects of modern **information retrieval**.

This book is an effort to partially fulfil this gap and should be useful for a first course on information retrieval as for a graduate course on the topic.

Sentence in Context

Thus, despite the fact that several people contributed to the text, this book is really much more a textbook than an edited collection of chapters written by separate authors. Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors to present a cohesive view of all the important aspects of modern information retrieval. Full-text of Chapters 1 (Introduction) and 10 (User Interfaces and Visualization) are available on-line, as well as a table of contents, exercises and resources for other chapters.

7 Modern **Information Retrieval**
www.sims.berkeley.edu/~hearst/irbook/

◁9 **Information Retrieval** (IR) has grown considerably in the last years with the expansion of the Web (World Wide Web) and the advent of modern and inexpensive graphic user interfaces and mass storage devices.

Figure 10.1. Necessary actions for relevance path traversal.

Subjects can visit the source document of any document representation by clicking its textual content. To see the next step in the relevance path they must *click* arrows next to representations (e.g., click arrow next to top-ranking sentence to highlight source document) or *hover* over representations (e.g., hover over title to see summary).

Since the Recommendation and Automatic systems used the movement of the mouse pointer over parts of the interface as an indication of relevance a timing mechanism was implemented to ensure these 'hovers' were intentional. That is, a searcher would have to remain over a document title for two seconds before the pop-up summary window appeared. In the studies in Chapter Four I demonstrated that a timing mechanism can be useful to tackle problems caused by accidental mouseovers in feedback systems that use implicit feedback techniques. Also, when the document summary appears, the other information in the background of the interface darkens and is disabled to ensure that it does not interfere with the examination of the summary and cannot be clicked accidentally.

The Recommendation and Automatic systems used the information that subjects interacted with as implicit feedback of their interests. The systems used this feedback to build a richer body of evidence and choose query terms to represent the information interacted with. In Table 10.2 I show the actions necessary for these systems to identify what is of interest to searchers; the indications in bold are those that comprise a relevance path. Providing the bolded indications in order, from top-ranking sentence to sentence in context means a searcher will traverse a complete relevance path.

Table 10.2
Implicit relevance indications.

Document Representation	Indication	Interpretation
Top-Ranking Sentence (TRS)	1. Click TRS	View document
	2. ***Click arrow on TRS***	Highlight document title
	3. Click '…' [35] at end of TRS	View remainder of sentence
Title	1. ***Hover for over two seconds***	View summary
	2. Click title	View document
Summary	1. Click text	View document
	2. ***Click arrow on Summary***	View sentence in context
Summary sentence	1. Click text	View document
	2. ***Click arrow on Summary***	View sentence in context
Sentence in context	1. ***Click text***	View document

A simple governing interaction model is that interacting with a document representation in any way is interpreted as a positive relevance indication. It can be seen in Table 10.2 that subjects can view the source document of a representation simply by clicking on its textual content. The mouse pointer changes when over these representations to indicate that they can be clicked. Also, all interaction with the document summary is regarded as an indication of interest. That is, all clicks in the summary are an indication of relevance for the text in the summary.

10.3.1.3 Term Selection Model

All systems use the term selection model chosen from the probabilistic implicit feedback framework to select query modification terms. As described later in this chapter they differ in how these terms are subsequently used.

10.3.1.4 Document/Sentence Reordering

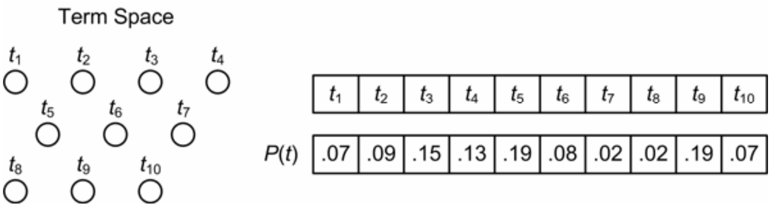
Two of the four possible retrieval strategies available for selection by the system or the subject involve reordering the most relevant documents and Top-Ranking Sentences. In my approach sentences are synonymous with small documents and the same approach is used to reorder documents and sentences. For consistency, in this section the term 'document' is synonymous with 'sentence'.

[35] To avoid unnecessary interface clutter, only the first 250 characters of a top-ranking sentence are shown at the interface. Ellipses are shown at the end of sentences where more text is available. Clicking on these ellipses shows the remainder of the sentence in a small area next to the mouse pointer. This is also used as an indication of interest.

The systems use a variation of the *tf.idf* approach to reorder the documents with respect to the query terms they contain. The inverse document frequency (*idf*) is regarded as a measure of importance of the term in the collection. In the approach used here, the values of the $P(t)$ assigned to terms in the term space can also be regarded as a measure of importance and the values are used instead of *idf* in this reordering. Unlike *idf* values, the $P(t)$ values alter to reflect the changing importance of the terms during a search. I now present an example of how this approach is used to rank documents or sentences.

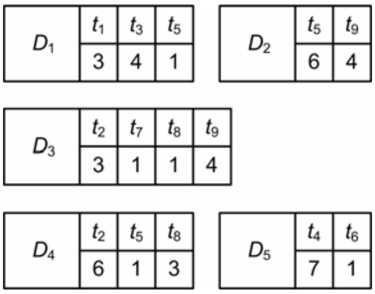
Example 10.1: Document Reordering

In this example there are five documents (D_1, D_2, D_3, D_4 and D_5) and the term space is in the same state as at the end of Example 7.1 (Chapter Seven). There are ten terms in the term space and the query contains terms t_2, t_5, t_8 and t_9. The weights assigned to each term in the term space are:



	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
$P(t)$	.07	.09	.15	.13	.19	.08	.02	.02	.19	.07

These weights are not revised during the reordering, but may change during the search, as a result of searcher interaction. They are used in conjunction with the frequency of terms within documents to produce a retrieval status value (RSV) used to rank documents. The term frequencies for each of the five documents in this example are:



D_1	t_1	t_3	t_5
	3	4	1

D_2	t_5	t_9
	6	4

D_3	t_2	t_7	t_8	t_9
	3	1	1	4

D_4	t_2	t_5	t_8
	6	1	3

D_5	t_4	t_6
	7	1

The documents are then ranked based on the scores of terms that reside in them and queries:

$$\boldsymbol{D_1} \quad \{t_2, t_5, t_8, t_9\} \cap \{t_1, t_3, t_5\} = \{t_5\}$$

$$\Rightarrow (.15 \times 1) = \mathbf{0.15}$$

$$\boldsymbol{D_2} \quad \{ t_2, t_5, t_8, t_9\} \cap \{t_5, t_9\} = \{t_5, t_9\}$$

$$\Rightarrow (.19 \times 6) + (.19 \times 4) = \mathbf{1.90}$$

$$\boldsymbol{D_3} \quad \{t_2, t_5, t_8, t_9\} \cap \{ t_2, t_7, t_8, t_9\} = \{t_2, t_8, t_9\}$$

$$\Rightarrow (.09 \times 3) + (.02 \times 1) + (.19 \times 4) = \mathbf{1.05}$$

$$\boldsymbol{D_4} \quad \{t_2, t_5, t_8, t_9\} \cap \{ t_2, t_5, t_8\} = \{t_2, t_5, t_8\}$$

$$\Rightarrow (.09 \times 6) + (.19 \times 1) + (.19 \times 3) = \mathbf{1.30}$$

$$\boldsymbol{D_5} \quad \{t_2, t_5, t_8, t_9\} \cap \{t_4, t_6\} = \varnothing$$

$$\Rightarrow \mathbf{0}$$

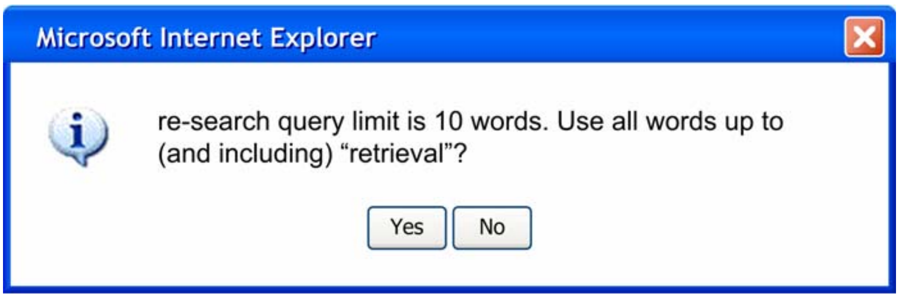
The document order based on the RSV is therefore D_2, D_4, D_3, D_1 and D_5. The documents that contained the query terms were ranked above those without. The top-ranked document (D_2) was ranked highly because it contained a large number of query terms that were regarded as important (i.e., had a high $P(t)$ value). It is conceivable that there could be a different document order if the searcher had interacted with different information before this action.

10.3.1.5 Initial Query Input and Restrictions on Length

The same initial query input screen is used by all experimental systems. This is the part of the system where the search typically begins. The look and feel of this initial interface is intentionally simple and contains a text input box, a submit button and access (through a link) to some details on the query syntax supported by the systems and the automatic exclusion of stopwords. Turtle (1994) found that searchers with no training in query formulation can experience difficulties in generating sound queries. He showed that unstructured queries containing only queries to separate the terms are more effective for searchers. Queries with embedded operators such as *, -, $ and + are meant to offer searchers greater control in query formulation. However, searchers may have difficulty using these operators because they are not consistent between search systems (Shneiderman *et al.*, 1997).

To prevent possible bias caused by previous search experience experimental subjects were not told that the systems were interfaces to Google. Queries were restricted to lists of terms separated by spaces and were automatically combined by the search engine. Queries submitted to Google have term order sensitivity (Muramatsu and Pratt, 2001). The system uses term proximity and exact phrase matching to give documents where terms that occur in the same order as the query and close proximity a higher weight. The concatenation of terms to form search phrases using “” was permitted. The use of search engine specific syntax such as ‘site:’ and ‘link:’ was discouraged.

Due to restrictions imposed by Google queries could not be longer than ten words. If the subject tried to submit a query of more than 10 words to any experimental system they were presented with an error message as shown in Figure 10.2.



re-search query limit is 10 words. Use all words up to (and including) “retrieval”?

Figure 10.2. Query length notification message.

The query is truncated at the tenth word but before doing so the searcher is asked if they want to proceed. In Figure 10.2 the tenth word in the query is ‘retrieval’ and all words that follow this will be ignored by the search system.

10.3.1.6 Reversal of Retrieval Strategies

In his book ‘Designing the User Interface’, Shneiderman (1998) stresses the importance of allowing users to reverse the effects of their interaction. In each experimental system the subject has the option to reverse the effect of any search decision made by them or by the system. This is done using a clickable ‘undo’ button shown in Figure 10.3.

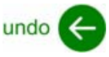


Figure 10.3. Retrieval strategy reversal (‘undo’) button.

The button intentionally resembles the ‘back’ button in Internet Explorer, the browser used for these experiments. Although the functionally was different (i.e., it does not take subjects back to the previous Web document), the underlying intent is similar (i.e., to reverse the last action). I assume that clicking this button is an indication of dissatisfaction with the outcome last search decision. The underlying implicit feedback framework does not consider such negative feedback, only positive indications are used. However, it is plausible that the reversal of decisions and the traversal of short relevance paths (without visiting the source document) could be used as an indication of disinterest and to lessen the weight of terms in those representations and modify the decision boundaries used when selecting new retrieval strategies.

10.3.1.7 Notification of Actions

The Recommendation and Automatic systems select new query terms and make search decisions for the subject as they search. They notify them of this by displaying messages in the bottom left-hand corner of the interface. However, if the searcher is looking at information in a different part of the screen they may be unaware a retrieval strategy has occurred or been recommended to them. To be sure they notice these actions the systems place an 'idea bulb' next to the mouse pointer when a strategy is followed or a recommendation is made. This is shown in Figure 10.4.

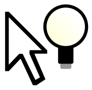


Figure 10.4. The 'idea bulb' notification at appears next to the mouse pointer (pictured).

This bulb disappears when the subject interacts with the suggested terms or notification messages in any way. Since the mouse pointer is the primary means of interacting with the search interface, communicating decisions via the pointer notifies searchers, but does not intrude on their search (i.e., they can simply ignore the bulb). The idea bulb supplements the Recommendation and Automatic system notifications, which appear in one part of the interface and may not be immediately noticeable if searcher attention is elsewhere. In this section I have described the similarities between the three experimental systems. In the next section I outline the differences between the experimental systems.

10.3.2 Differences

The differences between systems were necessitated by the hypotheses tested in this experiment. More specifically, the systems vary subject control over three main classes of decisions: selecting query terms, indicating relevance and making new search decisions (i.e., choosing retrieval strategies). In this section I describe the differences between systems in a set of pair-wise comparisons.

10.3.2.1 Checkbox and Recommendation

There are three differences between these systems; how new queries are created, how search decisions are made and how relevance information is communicated. The Checkbox system awaits the searcher's instruction and selects query terms that describe the information the searcher has explicitly marked as relevant. The searcher can add or remove their query terms. The Recommendation system does not require such direct indications and presents a list of potentially useful terms that can be added to the initial query. In both systems the subject has complete control over when a search decision is made and which decision is made. The Recommendation system recommends the retrieval strategy to the searcher, based on the estimated amount of information need change. The searcher has the option on whether to accept this recommendation.

10.3.2.2 Checkbox and Automatic

The differences between these systems lie in how search decisions are controlled and how relevance indications are provided. Retrieval strategies are controlled by the subject in the Checkbox system and by the information need tracking component in the Automatic system. Relevance is communicated directly (explicitly) in the Checkbox system and indirectly (implicitly) in the Automatic system.

10.3.2.3 Recommendation and Automatic

These experimental systems differ in how terms are selected for query modification and how search decisions are controlled. The Automatic system chooses terms and retrieval strategies on the subject's behalf. In contrast, the Recommendation system recommends terms and strategies.

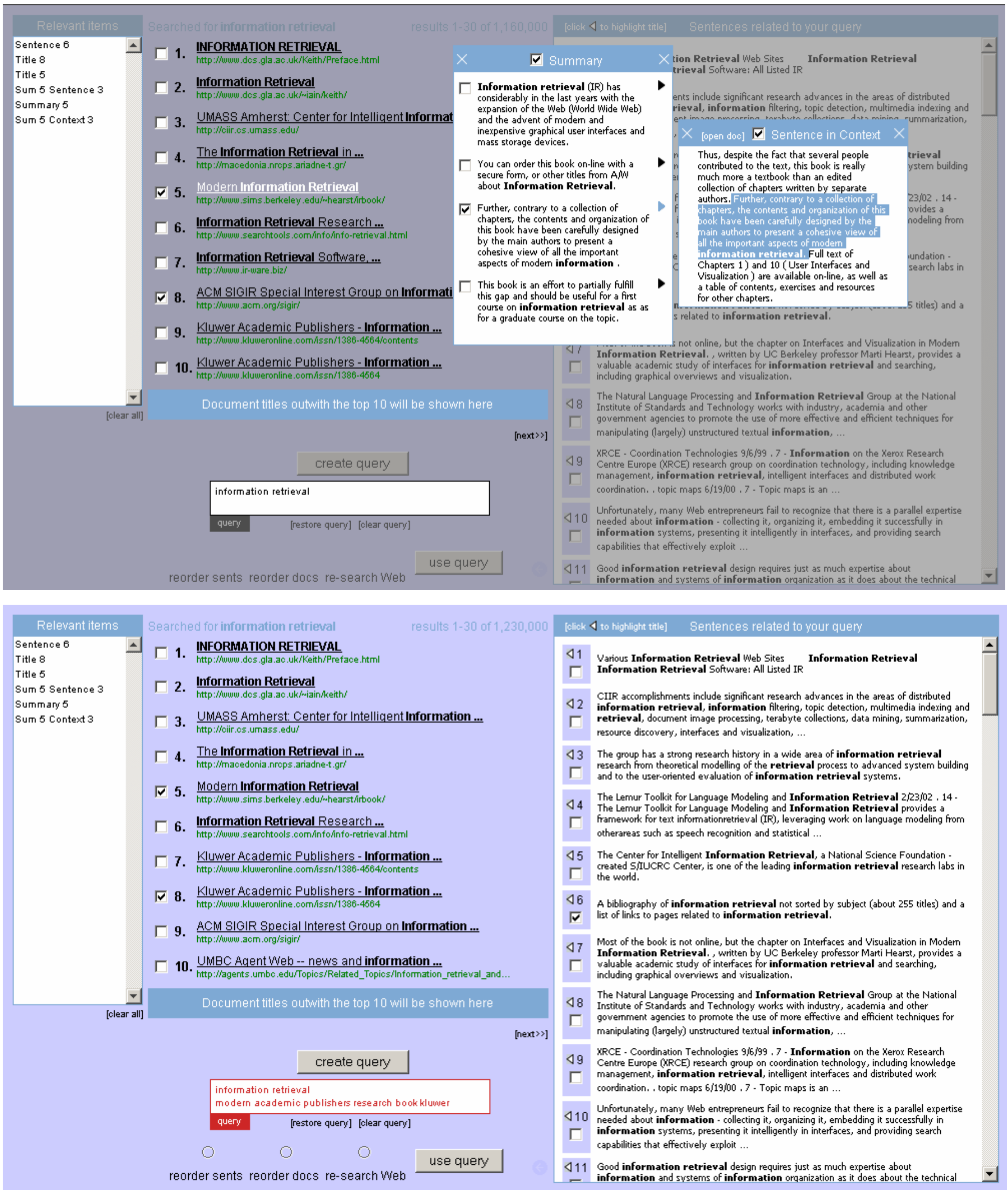
Overall, the systems differ in the amount of control they offer to the searcher. With additional control there is also extra responsibility for making query modification decisions and choosing appropriate retrieval strategies. In the Checkbox system there is also the additional burden of explicitly marking document representations. Beaulieu and Jones (1998) showed that such additional control is not always preferred by searchers and places additional demands on their finite cognitive resources. However, these systems allow searchers to indicate what information has relevant properties and may be more accurate than systems without this burden. In the next section the experimental systems are described in more detail.

10.4 Systems

The experimental systems each consist of an interface connected to Google with the architecture defined in Section 10.3.1.1. In this section I describe each of the three systems.

10.4.1 Checkbox System

This system allows subjects to communicate directly which document representations are relevant. A checkbox is shown next to each representation and the subject can choose which representations to mark. Marking a representation is an indication that its contents are relevant. The interface for this system is shown at two points during a search in Figure 10.6. The first part of the figure shows the summary window and sentence in context requested by the searcher. When 'Summary' or 'Sentence in Context' windows are requested the background darkens and is disabled to focus searcher attention on the active representation. Unlike the other experimental systems, all document representations in this system have checkboxes next to them that allow the searcher to mark them as relevant. In the second part of Figure 10.5 the searcher has requested assistance in creating a new query using the representations marked and extra terms have been added to the editable query entry box.



Relevant items
Sentence 6
Title 8
Title 5
Sum 5 Sentence 3
Summary 5
Sum 5 Context 3
[clear all]

Searched for **information retrieval** results 1-30 of 1,160,000

- □ 1. **INFORMATION RETRIEVAL** http://www.dcs.gla.ac.uk/Keith/Preface.html
- □ 2. **Information Retrieval** http://www.dcs.gla.ac.uk/~iain/keith/
- □ 3. UMASS Amherst: Center for Intelligent **Informat** http://ciir.cs.umass.edu/
- □ 4. The **Information Retrieval** in ... http://macedonia.nrcps.ariadne-t.gr/
- ☑ 5. Modern **Information Retrieval** http://www.sims.berkeley.edu/~hearst/irbook/
- □ 6. **Information Retrieval** Research ... http://www.searchtools.com/info/info-retrieval.html
- □ 7. **Information Retrieval** Software, ... http://www.ir-ware.biz/
- ☑ 8. ACM SIGIR Special Interest Group on **Informati** http://www.acm.org/sigir/
- □ 9. Kluwer Academic Publishers - **Information ...** http://www.kluweronline.com/issn/1386-4564/contents
- □ 10. Kluwer Academic Publishers - **Information ...** http://www.kluweronline.com/issn/1386-4564

Document titles outwith the top 10 will be shown here

[next>>]

create query

information retrieval

query [restore query] [clear query]

use query

reorder sents reorder docs re-search Web

☑ Summary

- □ **Information retrieval** (IR) has considerably in the last years with the expansion of the Web (World Wide Web) and the advent of modern and inexpensive graphical user interfaces and mass storage devices.
- □ You can order this book on-line with a secure form, or other titles from A/W about **Information Retrieval**.
- ☑ Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors to present a cohesive view of all the important aspects of modern **information** .
- □ This book is an effort to partially fulfill this gap and should be useful for a first course on **information retrieval** as as for a graphical course on the topic.

[open doc] ☑ Sentence in Context

Thus, despite the fact that several people contributed to the text, this book is really much more a textbook than an edited collection of chapters written by separate authors. Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors to present a cohesive view of all the important aspects of modern **information retrieval**. Full text of Chapters 1) and 10 (User Interfaces and Visualization) are available on-line, as well as a table of contents, exercises and resources for other chapters.

[click ◁ to highlight title] Sentences related to your query

- ◁7 □ s not online, but the chapter on Interfaces and Visualization in Modern **Information Retrieval**. , written by UC Berkeley professor Marti Hearst, provides a valuable academic study of interfaces for **information retrieval** and searching, including graphical overviews and visualization.
- ◁8 □ The Natural Language Processing and **Information Retrieval** Group at the National Institute of Standards and Technology works with industry, academia and other government agencies to promote the use of more effective and efficient techniques for manipulating (largely) unstructured textual **information**, ...
- ◁9 □ XRCE - Coordination Technologies 9/6/99 . 7 - **Information** on the Xerox Research Centre Europe (XRCE) research group on coordination technology, including knowledge management, **information retrieval**, intelligent interfaces and distributed work coordination. . topic maps 6/19/00 . 7 - Topic maps is an ...
- ◁10 □ Unfortunately, many Web entrepreneurs fail to recognize that there is a parallel expertise needed about **information** - collecting it, organizing it, embedding it successfully in **information** systems, presenting it intelligently in interfaces, and providing search capabilities that effectively exploit ...
- ◁11 Good **information retrieval** design requires just as much expertise about **information** and systems of **information** organization as it does about the technical

Relevant items
Sentence 6
Title 8
Title 5
Sum 5 Sentence 3
Summary 5
Sum 5 Context 3
[clear all]

Searched for **information retrieval** results 1-30 of 1,230,000

- □ 1. **INFORMATION RETRIEVAL** http://www.dcs.gla.ac.uk/Keith/Preface.html
- □ 2. **Information Retrieval** http://www.dcs.gla.ac.uk/~iain/keith/
- □ 3. UMASS Amherst: Center for Intelligent **Information ...** http://ciir.cs.umass.edu/
- □ 4. The **Information Retrieval** in ... http://macedonia.nrcps.ariadne-t.gr/
- ☑ 5. Modern **Information Retrieval** http://www.sims.berkeley.edu/~hearst/irbook/
- □ 6. **Information Retrieval** Research ... http://www.searchtools.com/info/info-retrieval.html
- □ 7. Kluwer Academic Publishers - **Information ...** http://www.kluweronline.com/issn/1386-4564/contents
- ☑ 8. Kluwer Academic Publishers - **Information ...** http://www.kluweronline.com/issn/1386-4564
- □ 9. ACM SIGIR Special Interest Group on **Information ...** http://www.acm.org/sigir/
- □ 10. UMBC Agent Web -- news and **information ...** http://agents.umbc.edu/Topics/Related_Topics/Information_retrieval_and...

Document titles outwith the top 10 will be shown here

[next>>]

create query

information retrieval
modern academic publishers research book kluwer

query [restore query] [clear query]

○ reorder sents ○ reorder docs ○ re-search Web

use query

[click ◁ to highlight title] Sentences related to your query

- ◁1 □ Various **Information Retrieval** Web Sites **Information Retrieval** **Information Retrieval** Software: All Listed IR
- ◁2 □ CIIR accomplishments include significant research advances in the areas of distributed **information retrieval**, **information** filtering, topic detection, multimedia indexing and **retrieval**, document image processing, terabyte collections, data mining, summarization, resource discovery, interfaces and visualization, ...
- ◁3 □ The group has a strong research history in a wide area of **information retrieval** research from theoretical modelling of the **retrieval** process to advanced system building and to the user-oriented evaluation of **information retrieval** systems.
- ◁4 □ The Lemur Toolkit for Language Modeling and **Information Retrieval** 2/23/02 . 14 - The Lemur Toolkit for Language Modeling and **Information Retrieval** provides a framework for text informationretrieval (IR), leveraging work on language modeling from otherareas such as speech recognition and statistical ...
- ◁5 □ The Center for Intelligent **Information Retrieval**, a National Science Foundation - created S/IUCRC Center, is one of the leading **information retrieval** research labs in the world.
- ◁6 ☑ A bibliography of **information retrieval** not sorted by subject (about 255 titles) and a list of links to pages related to **information retrieval**.
- ◁7 □ Most of the book is not online, but the chapter on Interfaces and Visualization in Modern **Information Retrieval**. , written by UC Berkeley professor Marti Hearst, provides a valuable academic study of interfaces for **information retrieval** and searching, including graphical overviews and visualization.
- ◁8 □ The Natural Language Processing and **Information Retrieval** Group at the National Institute of Standards and Technology works with industry, academia and other government agencies to promote the use of more effective and efficient techniques for manipulating (largely) unstructured textual **information**, ...
- ◁9 □ XRCE - Coordination Technologies 9/6/99 . 7 - **Information** on the Xerox Research Centre Europe (XRCE) research group on coordination technology, including knowledge management, **information retrieval**, intelligent interfaces and distributed work coordination. . topic maps 6/19/00 . 7 - Topic maps is an ...
- ◁10 □ Unfortunately, many Web entrepreneurs fail to recognize that there is a parallel expertise needed about **information** - collecting it, organizing it, embedding it successfully in **information** systems, presenting it intelligently in interfaces, and providing search capabilities that effectively exploit ...
- ◁11 Good **information retrieval** design requires just as much expertise about **information** and systems of **information** organization as it does about the technical

Figure 10.5. Checkbox system interfaces.

On the far left of the interface is a list of ‘Relevant items’ that describes which representations the subject has chosen so far. The nature of the interface, with pop-ups etc. is such that the subject may not see all representations they have marked relevant. This list allows them to keep track of what they have marked. In Figure 10.5 a number of document representations have been marked by the searcher. At any point the searcher can clear all representations they have marked or double-click an entry in the list of marked representations to highlight that particular representation. For clarity, from this point on all search interfaces are shown without the document summary and sentence in context pop-up.

The interface contains control options that allow the subject to request support with query formulation, modify the query and choose retrieval strategies. These options are shown in Figure 10.6.

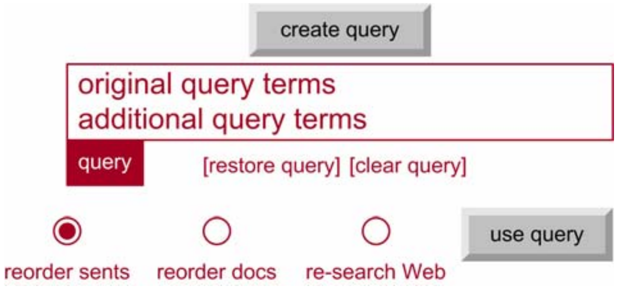


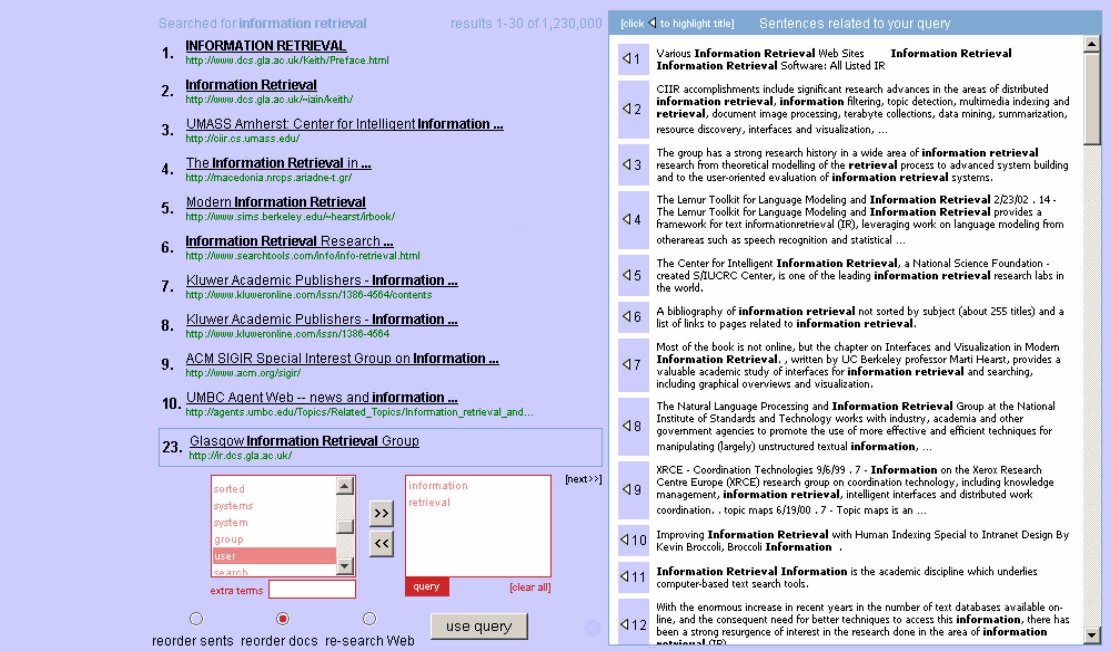
Figure 10.6. Term/retrieval strategy selection in the Checkbox system.

When they are satisfied with the document representations marked the subject can click the ‘create query’ button and a new query will be constructed. The presence of the button allows subjects to request assistance with query formulation. The term selection model treats each marked document representation as a separate relevance path and the order they were marked in is important. The terms chosen to expand the query are the six terms with the highest probability of relevance ($P(t)$ from Equation 7.10). These terms are appended onto the original query and presented in a search box for the searcher to edit, shown in Figure 10.6. The new query terms will be shown on a new line, below the original query.

In the Checkbox system the subject has control over the nature and timing of when search decisions are made. That is, at any time during their search they can choose the retrieval strategy (i.e., when to reorder the sentences, reorder the documents or re-search the Web) they feel is most appropriate.

10.4.2 Recommendation System

In the Recommendation system there are no checkboxes for the subject to explicitly mark what document representations are relevant. Instead, the system implicitly infers what is relevant from representations the subject has expressed an interest in through viewing or clicking. The search interface for the Recommendation system is shown in Figure 10.7.



1. **INFORMATION RETRIEVAL**
http://www.dcs.gla.ac.uk/Keith/Preface.html
2. **Information Retrieval**
http://www.dcs.gla.ac.uk/~iain/keith/
3. UMASS Amherst: Center for Intelligent **Information ...**
http://ciir.cs.umass.edu/
4. The **Information Retrieval** in **...**
http://macedonia.nrcps.ariadne-t.gr/
5. Modern **Information Retrieval**
http://www.sims.berkeley.edu/~hearst/irbook/
6. **Information Retrieval** Research **...**
http://www.searchtools.com/info/info-retrieval.html
7. Kluwer Academic Publishers - **Information ...**
http://www.kluweronline.com/issn/1386-4564/contents
8. Kluwer Academic Publishers - **Information ...**
http://www.kluweronline.com/issn/1386-4564
9. ACM SIGIR Special Interest Group on **Information ...**
http://www.acm.org/sigir/
10. UMBC Agent Web -- news and **information ...**
http://agents.umbc.edu/Topics/Related_Topics/Information_retrieval_and...

23. Glasgow **Information Retrieval** Group
http://ir.dcs.gla.ac.uk/

◁1 Various **Information Retrieval** Web Sites **Information Retrieval** **Information Retrieval** Software: All Listed IR

◁2 CIIR accomplishments include significant research advances in the areas of distributed **information retrieval**, **information** filtering, topic detection, multimedia indexing and **retrieval**, document image processing, terabyte collections, data mining, summarization, resource discovery, interfaces and visualization, ...

◁3 The group has a strong research history in a wide area of **information retrieval** research from theoretical modelling of the **retrieval** process to advanced system building and to the user-oriented evaluation of **information retrieval** systems.

◁4 The Lemur Toolkit for Language Modeling and **Information Retrieval** 2/23/02 . 14 - The Lemur Toolkit for Language Modeling and **Information Retrieval** provides a framework for text informationretrieval (IR), leveraging work on language modeling from otherareas such as speech recognition and statistical ...

◁5 The Center for Intelligent **Information Retrieval**, a National Science Foundation - created S/IUCRC Center, is one of the leading **information retrieval** research labs in the world.

◁6 A bibliography of **information retrieval** not sorted by subject (about 255 titles) and a list of links to pages related to **information retrieval**.

◁7 Most of the book is not online, but the chapter on Interfaces and Visualization in Modern **Information Retrieval**. , written by UC Berkeley professor Marti Hearst, provides a valuable academic study of interfaces for **information retrieval** and searching, including graphical overviews and visualization.

◁8 The Natural Language Processing and **Information Retrieval** Group at the National Institute of Standards and Technology works with industry, academia and other government agencies to promote the use of more effective and efficient techniques for manipulating (largely) unstructured textual **information**, ...

◁9 XRCE - Coordination Technologies 9/6/99 . 7 - **Information** on the Xerox Research Centre Europe (XRCE) research group on coordination technology, including knowledge management, **information retrieval**, intelligent interfaces and distributed work coordination. . topic maps 6/19/00 . 7 - Topic maps is an ...

◁10 Improving **Information Retrieval** with Human Indexing Special to Intranet Design By Kevin Broccoli, Broccoli **Information** .

◁11 **Information Retrieval Information** is the academic discipline which underlies computer-based text search tools.

◁12 With the enormous increase in recent years in the number of text databases available on-line, and the consequent need for better techniques to access this **information**, there has been a strong resurgence of interest in the research done in the area of **information retrieval** (IR)

Figure 10.7. Recommendation system interface.

At intervals of five [36] relevance paths, the system chooses a new set of potentially useful query terms and a retrieval strategy based on the level of change in its internal information need formulation since the last subject-controlled query submission. Terms are chosen that reflect the information viewed. The degree of change since the last time a new result set was generated is used to select the action the system will perform. The system chooses the top 20 most relevant terms and presents these in the ‘Recommended Terms’ box (Figure 10.8).

[36] This was chosen in pilot testing (including Pilot Test 1) and allowed the system to build a body of evidence sufficient to make decisions.

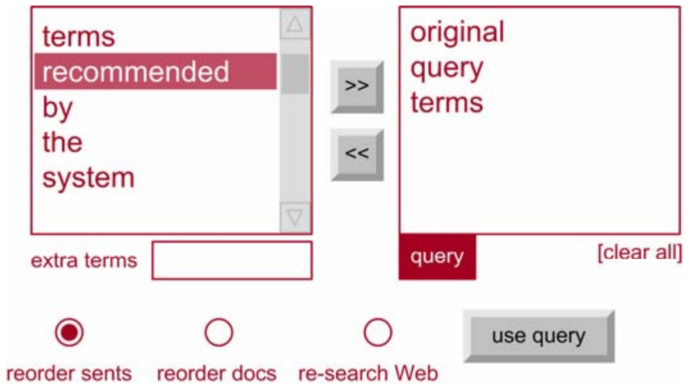


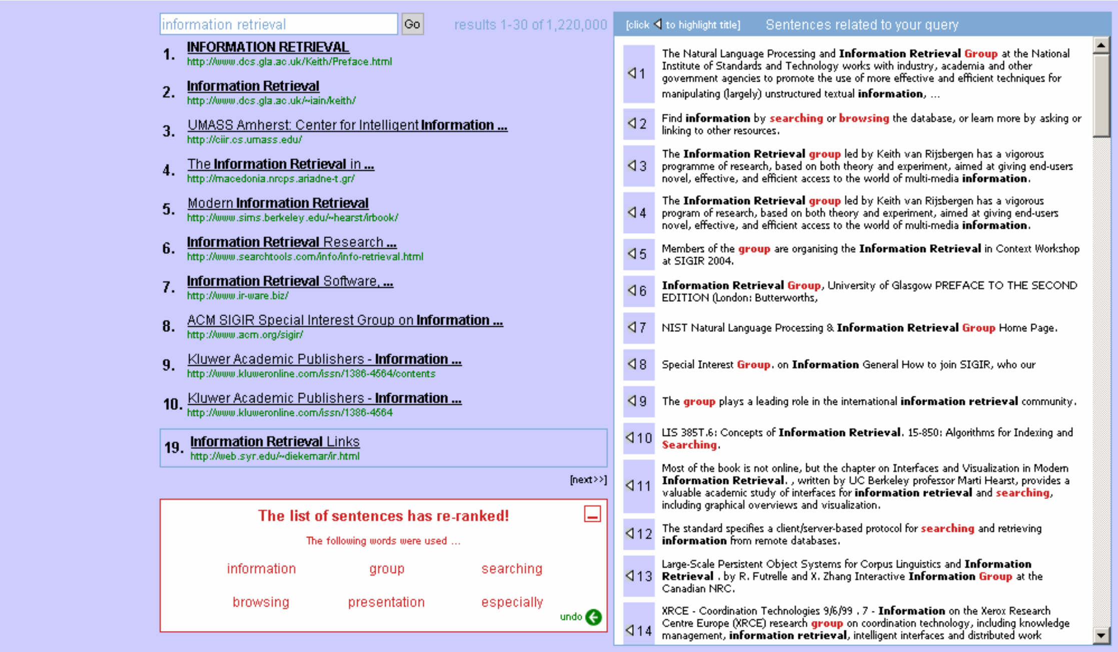
Figure 10.8. Term/retrieval strategy selection in the Recommendation system.

The subject can then control which terms are added to the query. Terms can also be deleted from the query. The '>>' and '<<' buttons can be used to transfer terms between the recommended list and the query. There is an 'extra terms' box where subjects can add additional terms to the query that are not in recommended terms list. When the subject clicks the '>>' button or presses 'enter' the term(s) in the box are added to the query. If the box contains more than one term the contents of it are tokenised and each token is added to the query separately. To reduce the number of erroneous terms that are transferred the searcher is only able to select and add one term at a time. Informal pilot testing of the interface revealed that subjects rarely want to add blocks of contiguous terms to the query at the same time. They preferred instead to be careful and selective about the terms they chose.

The system highlights the radio button for the retrieval strategy recommended by the experimental system. The subject does not have to agree with this recommendation and can choose another strategy or simply do nothing.

10.4.3 Automatic System

The Automatic system obtains its relevance assessments implicitly in the same way as the Recommendation system. However, the system retains control of the search decisions taken and the terms used. Rather than recommending terms and retrieval strategies, the Automatic system chooses them, without direct instruction. The interface is shown in Figure 10.9.



1. **INFORMATION RETRIEVAL**
http://www.dcs.gla.ac.uk/Keith/Preface.html
2. **Information Retrieval**
http://www.dcs.gla.ac.uk/~iain/keith/
3. UMASS Amherst: Center for Intelligent **Information ...**
http://ciir.cs.umass.edu/
4. The **Information Retrieval** in ...
http://macedonia.nrcps.ariadne-t.gr/
5. Modern **Information Retrieval**
http://www.sims.berkeley.edu/~hearst/irbook/
6. **Information Retrieval** Research ...
http://www.searchtools.com/info/info-retrieval.html
7. **Information Retrieval** Software, ...
http://www.ir-ware.biz/
8. ACM SIGIR Special Interest Group on **Information ...**
http://www.acm.org/sigir/
9. Kluwer Academic Publishers - **Information ...**
http://www.kluweronline.com/issn/1386-4564/contents
10. Kluwer Academic Publishers - **Information ...**
http://www.kluweronline.com/issn/1386-4564

19. **Information Retrieval** Links
http://web.syr.edu/~diekemar/ir.html

The list of sentences has re-ranked!

The following words were used ...

information group searching
browsing presentation especially

Sentences related to your query

◁1 The Natural Language Processing and **Information Retrieval Group** at the National Institute of Standards and Technology works with industry, academia and other government agencies to promote the use of more effective and efficient techniques for manipulating (largely) unstructured textual **information**, ...

◁2 Find **information** by **searching** or **browsing** the database, or learn more by asking or linking to other resources.

◁3 The **Information Retrieval group** led by Keith van Rijsbergen has a vigorous programme of research, based on both theory and experiment, aimed at giving end-users novel, effective, and efficient access to the world of multi-media **information**.

◁4 The **Information Retrieval group** led by Keith van Rijsbergen has a vigorous program of research, based on both theory and experiment, aimed at giving end-users novel, effective, and efficient access to the world of multi-media **information**.

◁5 Members of the **group** are organising the **Information Retrieval** in Context Workshop at SIGIR 2004.

◁6 **Information Retrieval Group**, University of Glasgow PREFACE TO THE SECOND EDITION (London: Butterworths,

◁7 NIST Natural Language Processing & **Information Retrieval Group** Home Page.

◁8 Special Interest **Group**. on **Information** General How to join SIGIR, who our

◁9 The **group** plays a leading role in the international **information retrieval** community.

◁10 LIS 385T.6: Concepts of **Information Retrieval**. 15-850: Algorithms for Indexing and **Searching**.

◁11 Most of the book is not online, but the chapter on Interfaces and Visualization in Modern **Information Retrieval**. , written by UC Berkeley professor Marti Hearst, provides a valuable academic study of interfaces for **information retrieval** and **searching**, including graphical overviews and visualization.

◁12 The standard specifies a client/server-based protocol for **searching** and retrieving **information** from remote databases.

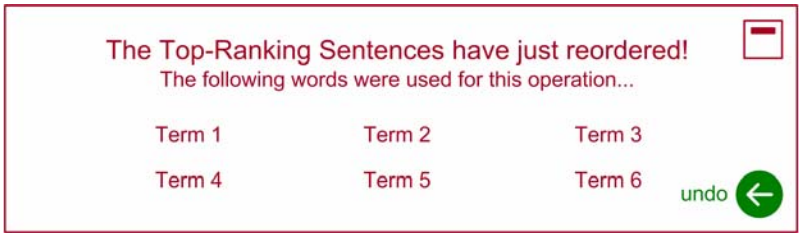
◁13 Large-Scale Persistent Object Systems for Corpus Linguistics and **Information Retrieval** . by R. Futrelle and X. Zhang Interactive **Information Group** at the Canadian NRC.

◁14 XRCE - Coordination Technologies 9/6/99 . 7 - **Information** on the Xerox Research Centre Europe (XRCE) research **group** on coordination technology, including knowledge management, **information retrieval**, intelligent interfaces and distributed work

Figure 10.9. Automatic system interface (with maximised notification, Figure 10.10).

This system allows the subject to edit their original query and retrieve a new set of documents. No provision is made for the subject to formulate a query for reordering sentences or documents, these actions are controlled by the system. The system chose terms automatically and acts on the subject's behalf. Since subjects could not control the terms that were used it was necessary for this system to be able to replace the original query terms. If the information need changed during the search, the presence of the original terms would have meant the system could not totally adapt to that change. As in the Checkbox and Recommendation systems the new query is limited to a maximum of 10 terms as this is the maximum number of query terms supported by the Google search engine.

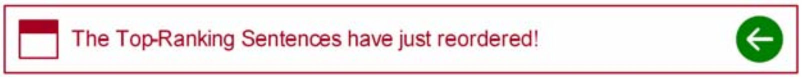
The system notified subjects that a new set of documents had been retrieved or the already retrieved information had been restructured using notifications at the search interface. These notifications were in two forms: maximised notification and minimised notification, shown in Figure 10.10 and Figure 10.11 respectively.



The Top-Ranking Sentences have just reordered!
The following words were used for this operation...

Term 1 Term 2 Term 3
Term 4 Term 5 Term 6

Figure 10.10. Maximised Automatic system notification.



The Top-Ranking Sentences have just reordered!

Figure 10.11. Minimised Automatic system notification.

The minimised notification is less intrusive, but is also less informative and does not tell the subject which terms are used. The subject can switch between the different forms of notification by clicking on the notification message.

10.5 Chapter Summary

Three experimental systems have been described this chapter. These systems were created to test the hypotheses given in Chapter Nine. The systems allow relevance information to be communicated in different ways, and for subjects to have varying degrees of control over how new queries are created and how search decisions are made during their search. All systems use the probabilistic implicit feedback framework described in Chapter Seven. In the next chapter the results of the experiment involving these systems are presented and analysed.

Chapter 11

Experimental Results and Analysis

11.1 Introduction

In this chapter the results of the user experiment described in the two preceding chapters are presented. The experiment tests three search interfaces that vary searcher control over interface decisions, and the probabilistic implicit feedback framework (from Chapter Seven) that underlies them. Experimental subjects attempted search scenarios on the experimental systems and provided feedback on their experience through questionnaires and comments made during informal discussions. I focus on results that relate to each of the three research hypotheses originally proposed at the end of Chapter Nine:

Interface support (Hypothesis 1)

> The interface support provided by the experimental systems was liked by subjects and facilitated effective information access.

Information need detection (Hypothesis 2)

> Subjects found the terms chosen by the probabilistic implicit feedback framework valuable and worthwhile.

Information need tracking (Hypothesis 3)

> Subjects found the retrieval strategies chosen by the probabilistic implicit feedback framework valuable and worthwhile.

The hypotheses are tested in terms of search effectiveness and subject preference. A total of 48 subjects, with different levels of search experience participated in the experiment. Subjects were classified into two groups – *inexperienced* and *experienced* – each containing 24 volunteers and a mixture of males and females. Results are presented for inter-system (Checkbox *versus* Recommendation *versus* Automatic) and inter-group (inexperienced *versus* experienced) comparisons. The significance of experimental results is tested at $p < .05$, unless otherwise stated. As in Chapter Ten S_{Check}, S_{Recomm} and S_{Auto} are used to denote the Checkbox, Recommendation and Automatic experimental systems respectively. In this chapter I also present results on the novel interface components (i.e., the relevance paths and increased information content at the search interface) and the search tasks.

The results presented in this chapter are based on questionnaire responses and system logs generated during interaction. The evidence is supported by informal subject feedback and my own observations. Questionnaires used five point Likert scales and semantic differentials with a lower score representing more agreement with the attitude object. The arrangement of positive (e.g., 'easy', 'relaxing') and negative (e.g., 'difficult', 'stressful') descriptors was randomised so that a positive assessment would be represented sometimes by a high score (i.e., approaching 5) and sometimes by a low one (i.e., approaching 1). This ensured that subjects applied due care and attention when completing the differentials (Busha and Harter, 1980). At the analysis stage the high positive scores are reversed so that in all cases the positive assessments were represented by low scores.

No assumptions are made about the normality of the data gathered during the experiment. Non-parametric statistical tests are used to test for statistical significance since these tests do not make any assumptions about the underlying distribution of the data. Also, since much of the data gathered was ordinal in nature (e.g., Likert scales and semantic differentials) these methods are more appropriate than their parametric equivalents. As described earlier, subjects were divided into two groups, *inexperienced* and *experienced*. The analysis presented involves *within-group* comparisons (e.g., one subject group with two or more systems) and *between-group* comparisons (e.g., comparing different subject groups on the same system). Where appropriate Dunn's *post hoc* tests (multiple comparison using rank sums) are applied to reduce the likelihood of Type I errors (i.e., rejecting null hypotheses that are true). The results across both subject groups are combined to form an 'Overall' group that gives a holistic view of the experimental findings for all subjects. The experimental design is a 2 × 3 factorial with *search experience* (2 levels) and the *experimental systems* (3 systems) as the main effects; tests are run for interaction between these where appropriate.

I begin this chapter by presenting results on the search process (Section 11.2) and the tasks attempted (Section 11.3). Tasks are analysed separately and relative to subject perceptions and measures of search effectiveness. This is followed by findings on the interface support (Hypothesis 1) (Section 11.4) and the terms and strategies selected by the probabilistic framework (Hypotheses 2 and 3) (Section 11.5 and 11.6 respectively). In Section 11.7 this chapter concludes with a summary of the experimental findings. This experiment was in part a study of searcher control in interactive information retrieval. As such, the findings presented in this chapter focus on subjective impressions of the interface support mechanisms the experimental systems offer.

11.2 Search Process

In this section I present results on the search subjects performed. Whilst this analysis is not necessary to test the hypotheses, the factors may have an impact on subject perceptions. Each subject was asked to describe various aspects of their experience on each experimental system. The results presented are from questionnaire and informal subject comments, both during the search and after the experiment. Subjects were asked about their search and the quality of the information retrieved by each of the experimental systems.

11.2.1 Perceptions of Search

Subjects were asked to complete four semantic differentials about their search: 'relaxing'/'stressful', 'interesting'/'boring', 'restful'/'tiring' and 'easy'/'difficult'. The average value in relation to each positive differential is shown in Table 11.1. The 'Overall' value is derived from all four differentials and shows how the process is perceived across all subjects. For each differential in each subject group, the most positive average differential response is shown in bold. Below Table 11.1 and for each table in this chapter I use *n* to represent the number of trials in each cell. For example, in the table there are 24 trials in each 'Inexperienced' cell, 24 trials in each 'Experienced' cell and 48 trials in each 'Overall' cell.

Table 11.1

Subject perceptions of the search process (range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
relaxing	2.75	2.33	**2.17**	2.67	2.25	**2.21**	2.71	**2.29**	2.19
interesting	2.70	2.54	**2.38**	2.08	**1.88**	2.21	2.40	**2.21**	2.30
restful	2.79	**2.71**	**2.71**	2.71	**2.25**	2.33	2.75	**2.48**	2.52
easy	2.75	**2.38**	2.67	2.58	**2.33**	2.50	2.67	**2.36**	2.59
all	2.75	2.49	2.48	2.51	2.18	2.31	2.63	2.34	2.40

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

A Friedman Rank Sum Test was run for each differential within each group. The test tries to answer the question: If the different systems really are identical, what is the chance that random sampling would result in sums of ranks as far apart (or more so) as observed? Since this analysis involved multiple comparisons, I use a Bonferroni correction to control the experiment-wise error rate and set the *alpha level* (α) to .0125 i.e., .05 divided by 4, the number of tests performed. This correction reduces the number of Type I errors i.e., rejecting null hypotheses that were true. The results showed significant differences for the 'relaxing', 'interesting' and 'easy' differentials (*inexperienced*: all $\chi^2(2) \geq 14.26$, all $p < .001$) and 'relaxing', 'interesting', 'restful' and 'easy' differentials (*experienced*: all $\chi^2(2) \geq 14.83$, all $p < .001$ and *overall*: all $\chi^2(2) \geq 16.22$, all $p < .001$). [37] A Dunn's *post hoc* test was applied for each system in each subject group and found that for those differentials all differences were significant. The Recommendation system generally created a more pleasant search experience than the other systems; the Checkbox system was generally worse. Subjects found searches in the Recommendation system more interesting than in the other systems. The interface support provided by the system may have enabled subjects to view a broader range of documents or more fully explore those that interested them rather than dedicating time to explicitly assessing relevance.

The analysis also revealed significant differences in the differentials between the subject groups for the 'interesting', 'restful' and 'easy' differentials with a Mann-Whitney Test (all $U(24) \geq 399$, $\alpha = .0125$, all $p \leq .011$). To test for interaction effects between the two main effects; search experience and experimental system, and the dependent variable (i.e., the differential value) I ran a Kruskal-Wallis Test for each differential using the technique described by Meddis (1984, pp. 305-313). The test tries to answer the question: If the populations really have the same median, what is the chance that random sampling would result in sums of ranks as far apart (or more so) as observed in this experiment? The test returns an *H*-statistic that can use the Chi-square test to determine its significance (Siegel and Castellan, 1988). The results showed that for all differentials there was no significant interaction between search experience and system ($\chi^2(2) = 2.10$, $p = .35$). This demonstrates that the influence of the main effects on one another was not sufficient to affect the conclusions I can draw about each of them. This approach will be used where appropriate to test for interaction effects during this chapter.

11.2.2 Information value

The quality of information retrieved by search systems may have affected subject perceptions of them and could therefore influence the results described later in this chapter. To measure the quality of the information retrieved by the experimental systems throughout the search

[37] For large sample sizes the critical values of the Chi-squared distribution can be used to determine the statistical significance of the Friedman Rank Sum Test (Siegel and Castellan, 1988). Chi-Squared tests are represented by the notation χ^2(*degrees of freedom*).

subjects were asked for their opinion. On a Likert scale subjects indicated the extent to which they agreed with the attitude statement: *I think there was better information available (that the system did not help me find).* The average responses, for different systems and different subject groups are shown in Table 11.2.

Table 11.2

Quality of information retrieved by the experimental systems (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
3.00	**2.92**	2.96	3.08	**3.04**	2.96	3.04	**2.98**	2.96

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Subjects commented that they did not notice much difference between the quality of the information returned by the experimental systems. Since all systems use the same retrieval architecture the results retrieved may be very similar (and for the same query identical). The techniques presented in this thesis encourage interaction with the top-ranked document set. Whilst the systems offer different interface support mechanisms (necessitated by the experimental hypotheses) they use the same underlying retrieval techniques and retrieve the same documents in response to the same queries. Friedman Rank Sum Tests were used within each subject group to test for statistically significant differences; none were significant (*inexperienced*: $\chi^2(2)$ = 2.34, p = .310; *experienced*: $\chi^2(2)$ = 2.55, p = .280; *overall*: $\chi^2(2)$ = 2.53, p = .282). The difference between subject groups was not significant (*U*(24) = 305, p = .36) and there were no interaction effects between systems and search experience ($\chi^2(2)$ = .89, p = .64) This suggests that subjects did not notice a difference in the quality of the information retrieved between systems, and this is therefore unlikely to contribute to any inter-system differences reported later in this chapter. In the next section results obtained on tasks and task categories are presented and analysed.

11.3 Tasks

As suggested in Chapter Two, the experimental search task can have a large effect on an experiment. In this section the results on the tasks attempted are presented and analysed to discern whether the tasks had an effect on subject perceptions of the experimental systems and interaction with them. Subjects were able to choose tasks from six search topics in three task categories, one task per category. In this section I analyse the reasons subjects gave for their choice, the nature of the tasks they chose and other subject perceptions. Where appropriate, I analyse the results on a per task category (i.e., pre-focus, focus-formation and post-focus) and per system basis. The results presented in this section are not directly associated with any of the three experimental hypotheses but provide interesting insight into the experiment nonetheless.

11.3.1 Selection

The experimental design allowed subjects to choose the topic of their first search task from six options, their second topic from five options, and their third from four. [38] I was interested in *why* subjects had chosen their tasks as this may help explain anomalous findings and provide insight beneficial for the development of search tasks in future work. That is, if one can establish why subjects chose search tasks these criteria can be used to create similar tasks in the future. On the 'Search' questionnaire subjects were offered six possible explanations for their choice of task: 'interest', 'familiarity', 'no doable alternatives', 'least boring', 'no reason' and 'other'. They were asked to choose the reason that best described the rationale behind their task selection. The divided bar in Figure 11.1 illustrates the reasons given by subjects for choosing tasks.

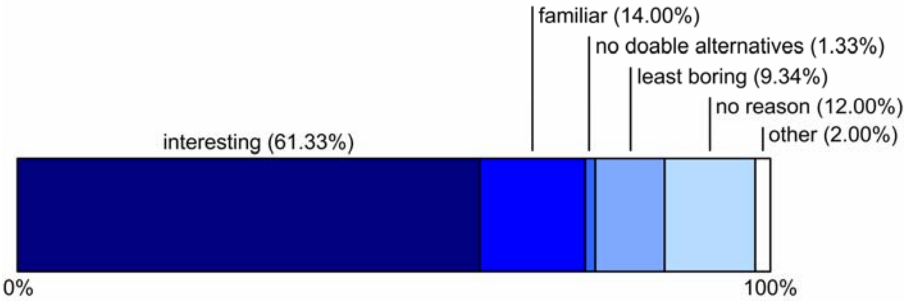


Figure 11.1. Reasons given by subjects for choosing search tasks.

The level of interest in the topic of the task appears to be the major contributory factor in deciding whether to choose a task from a number of alternatives. This supports the findings of Pilot Test 1 and the suggestion made by Borlund (2000b) that when creating tasks for interactive experimentation it is important to capture the interest of experimental subjects.

11.3.2 Nature

In this section I analyse the nature of the search tasks through subject perceptions of them generally, their perceptions of task success and the clarity of the information need created by the search tasks.

[38] Due to potential learning effects, subjects were not permitted to choose the same search topic for more than one search task.

11.3.2.1 Clarity and Complexity

Search tasks can influence subject perceptions of an experimental system or the entire experiment. For this reason it was important to determine if there were any expected or unexpected differences between tasks. Differences in the clarity and complexity of tasks between task groups were expected, since this was varied as part of the experimental design. Subjects were asked to indicate on semantic differentials how ‘clear’/‘unclear’ and ‘simple’/‘complex’ the tasks were. The average differential responses are shown in Table 11.3 for each task category and system type.

Table 11.3
Task characteristics across categories and experimental systems (range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus
clear	3.12	2.75	**2.31**	2.96	2.80	**2.36**	3.04	2.78	**2.34**
simple	2.87	2.54	**2.01**	2.72	2.40	**1.95**	2.80	2.47	**1.98**
all (*task*)	3.00	2.65	2.16	2.84	2.60	2.16	2.92	2.63	2.16
	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
clear	1.54	1.55	**1.48**	**1.33**	**1.33**	1.37	1.44	1.44	**1.43**
simple	2.08	2.00	**1.98**	**1.92**	1.93	2.00	2.00	**1.92**	1.96
all (*system*)	1.81	1.78	1.73	1.63	1.63	1.83	1.69	1.68	1.70

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

Pilot Test 2, described in Chapter Nine (Section 9.2.2), tested the clarity and simplicity of the tasks prior to the experiment. The pilot test showed that the tasks were all of similar levels and therefore unlikely to introduce unwanted task effects. However, it is perhaps more important to test how subjects perceived the tasks during the experiment as external factors may influence their perceptions. Table 11.3 also presents subject perceptions of the search task for different task categories and different systems. Since all tasks were created independent of the system I would expect no significant relationship between the task and system. This was verified by a Friedman Rank Sum Test applied to each differential in each subject group (all $\chi^2(2) \leq 2.41$, all $p \geq .30$).

The tasks were meant to simulate information needs at different stages in the information seeking process (ISP) and encourage different information seeking behaviours (Kuhlthau, 1991). The tasks were developed using the framework proposed by Bell and Ruthven (2004) and the complexity of the search tasks was varied as part of the experimental design. Therefore, subject perceptions of task complexity were important. The tasks were designed in such a way that the *pre-focus* task was designed to be more complex than the *focus formation* task, which was in turn designed to be more complex than the *post-focus* task. If this was implemented successfully, I would expect a drop in the differential value for clarity and simplicity from left to right within each subject group in Table 11.3; this was generally the case. The *pre-focus* tasks were vague and required information from multiple sources. Subjects found these tasks difficult and classified tasks in this category as least 'clear' and 'simple'. The *post-focus* tasks provided subjects with more information to use to begin and conduct their search. Subjects generally found tasks in this category the more 'clear' and 'simple' than those from other categories. These findings were significant with a series of Friedman Rank Sum Tests (all $\chi^2(2) \geq 7.73$, all $p \leq .021$). Overall, the categorisation of tasks appears to concord with general subject perceptions of their clarity and simplicity. There were no significant differences between subject groups (Mann-Whitney Test, all $U(24) \leq 318$, $\alpha = .0167$, all $p \geq .24$) and no significant interaction effects between search experience and task categories (all $\chi^2(2) \leq 1.43$, all $p \geq .49$). However, there are interaction effects between search experience and systems for both differentials (*clear*: $\chi^2(2) = 1.31$, $p = .52$, *simple*: $\chi^2(2) = 1.31$, $p = .52$). The experimental systems appear to affect subject perceptions of clarity and simplicity of the search task; this affects both subject groups differently. Inexperienced subjects found searches on the Automatic system more clear and simple, perhaps because it helped them more directly. In contrast, experienced subjects found searches on the Checkbox and Recommendation systems more clear and simple, perhaps because it gave them control.

To develop a more complete picture of task effects the 'Search' questionnaire contained further questions on task success and information need clarity. I now present findings on each of these questions.

11.3.2.2 Task Success

Subject perceptions of task success are important since search systems are designed to help searchers satisfy their information needs and their desire to complete the search task they are undertaking. Also, since simulated work tasks situations were used to encourage personal relevance assessments, it is only searchers who can truly judge whether a task is complete. After each search task, subjects were asked to indicate on a five point Likert scale the extent to which they agreed with the statement *I believe I have succeeded in my performance of this task*. In Table 11.4 I present subject perceptions of task success, averaged across different groups of experimental subjects.

Table 11.4

Subject perceptions of task success (range 1-5, lower = better).

Scale	Inexperienced			Experienced			Overall		
	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
Task success	2.43	**2.23**	2.46	2.50	**2.39**	2.41	2.42	**2.31**	2.44

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Friedman Rank Sum Tests were applied between systems on the same subject groups. For experienced subjects there were no significant inter-system differences ($\chi^2(2)$ = 3.67, p = .160). However, the inter-system differences for the inexperienced subjects appeared significant ($\chi^2(2)$ = 8.54, p = .014) suggesting that for this group at least one of the experimental treatments (systems) differed from the rest. The application of Dunn's *post hoc* tests revealed significant differences between the Recommendation system and the Checkbox/Automatic systems (all Z ≥ 2.01, all p ≤ .022). Other comparisons did not reveal significant differences. The Recommendation system appears to help inexperienced subjects complete search tasks. There were no significant differences between subject groups (Mann-Whitney Test, all *U*(24) ≤ 322, all p ≥ .24) and no significant interaction effects between search experience and systems ($\chi^2(2)$ =.70, p = .71).

11.3.2.3 Information Need Clarity

Each subject attempted tasks from three task categories – pre-focus, focus formation and post-focus. The tasks varied in complexity, with different categories requiring information from different numbers of sources and different types of information. In Table 11.5 I present the average five point Likert scale response to the attitude statement: *I had an exact idea of the type of information I wanted.*

Table 11.5

Subject awareness of information required (range 1-5, lower = better).

Scale	Inexperienced			Experienced			Overall		
	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus
Awareness	2.87	2.60	**2.10**	2.54	2.12	**1.94**	2.71	2.36	**2.02**

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

As task complexity increased, subject awareness of the information required decreases. An effect of this may be that subjects are less able to choose query terms and make search decisions, and therefore need more support from the search system. Mann-Whitney Tests were applied between the independent subject groups. The results revealed significant differences for *pre-focus* (*U*(24) = 399, p = .011), *focus-formation* (*U*(24) = 405, p < .001) and *post-focus* (*U*(24) = 396, p = .013) task categories. Experienced subjects appeared more aware of the type of information required during search tasks in each task category. Their enhanced search experience may mean that these subjects are better able to identify what information is necessary to complete their search.

11.3.3 Task Preference

Subjects attempted a task on each of the three systems. Afterwards they were asked to rank the tasks in their order of preference. No instructions were given on what factors to base their decision on, but subjects were asked to explain their ordering. The average subject rank for each task category is shown in the Table 11.6 for each subject group and across all subjects.

Table 11.6
Subjects' preferred task rank order (range 1-3, lower = better).

Inexperienced			Experienced			Overall		
Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus
2.25	2.00	**1.79**	2.21	1.92	**1.92**	2.23	1.96	**1.85**

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

A Kruskal-Wallis test was applied to the rankings in each subject group, and overall across all subjects. The results showed significant differences in the rankings assigned by inexperienced subjects ($\chi^2(2) = 11.04$, p = .004), experienced subjects ($\chi^2(2) = 8.85$, p = .012) and overall ($\chi^2(2) = 10.23$, p = .006). Dunn's *post hoc* tests were used to compare the task categories within each group. There were significant differences in the inexperienced group between all category pairs and in the experienced group between all pairs except *focus formation* and *post-focus* (all *Z* = 1.23, p = .109). Experienced subjects preferred the two less complex tasks but there was no discernable difference in the ranking between them.

I also compared the task preference between inexperienced and experienced subject groups. A Mann-Whitney Test was applied between the groups to determine the significance of any differences. The results showed that the rankings did not differ significantly (*U*(24) = 347, p = .112). That is, there was no discernable difference in the type of task inexperienced and experienced subjects prefer.

Subjects were asked to provide an explanation for their ranking. A variety of explanations were offered, however the most popular, in descending order of frequency were: 'interest in tasks', 'easiness of tasks', 'familiarity with similar tasks', 'task complexity', 'experimental systems' and 'task completion'. Subjects appear to place importance on the factors that influence their ability to complete search tasks.

A deeper examination of the subject comments revealed a split between the three task categories. That is, subjects appeared to notice differences between the categories and *how* the categories differed (i.e., in complexity). Since subjects were not informed that the tasks were categorised in this way, they are making their own inferences and seem able to discern even subtle variations in task complexity. In Table 11.7 examples of the comments made by subjects are provided.

Table 11.7
Subject comments on task categories
(numbers in brackets reflect the concept/statement frequency).

Pre-focus	Focus-formation	Post-focus
1. "research-based"	1. "more focused" (2)	1. "knew what to expect"
2. "complex" (2)	2. "hard to make initial query"	2. "clear" (3)
3. "very loose"	3. "specific topic"	3. "more technical"
4. "not very specific"		4. "easy to make initial query"
5. "hard to make initial query"		5. "precise information" (2)
6. "required further interaction"		6. "know exactly what I looked for"
7. "didn't know where to look"		7. "specific topic"
8. "open subject"		8. "more effective for queries"
9. "hard to find exact information"		

n = 48

As can be seen from the selection of comments, subjects appeared able to determine that tasks in the three categories differed in complexity. The difference between the comments in the *pre-* and *post-focus* categories is more apparent than other pair-wise differences. Subjects were not asked specifically about the nature of the task so not all subjects provided feedback of this kind. Others chose to make reference to the information retrieved by the experimental system, their own previous search experiences and task specifics (e.g., one subject chose to write "did you know there are 18,000 dust mites in one gram of dust?").

In this section the search process and search tasks attempted by subjects have been analysed. Since these factors affect subject perceptions of the experimental systems and the experiment as a whole it is important to consider them in an analysis such as this. The search tasks play a vital role in facilitating interaction with the search systems. Therefore, it was important to establish why tasks were chosen and whether the task categories were interpreted by subjects as they were meant to be (i.e., whether the level of task complexity as perceived by subjects matched that intended in the task categorisation). The findings presented in this section demonstrate that the Recommendation system leads to a more pleasant search and subject perceptions match the task categorisation. In what follows in this chapter I present and analyse results related to each of the three experimental hypotheses. In the next section I begin with the first, interface support.

11.4 Hypothesis 1: Interface Support

This section presents results related to the first experimental hypothesis: *the interface support provided by the experimental systems was liked by subjects and facilitated effective information access*. This hypothesis was divided into a number of sub-hypotheses that are tested in this section. To test these I analyse results obtained from a combination of questionnaire responses, system logs, informal subject comments, and my own observations. The interface support provided by all three experimental systems is compared based on how new queries are constructed, how retrieval strategies are chosen, how relevance information is conveyed and how (where appropriate) the system notified the subject of decisions it makes. The main differences between the three experimental systems are in the control they give subjects over aspects of their search.

11.4.1 Relevance Paths and Content

All systems present a large amount of information at, what I have referred to as, 'content-rich' search interfaces. Subjects were asked to express their opinion of this content in the 'Search' questionnaire and informally at the end of the experiment. As there are no path and content differences *between* systems, I only compare results between subject groups (i.e., inexperienced *versus* experienced).

From observations and informal post-search interviews, subjects appeared to use the relevance paths and found the increased levels of content shown at the search interface of value in their search. This is important, as the success of the both systems – especially the Recommendation and Automatic systems – is dependent on using these interface components. All experimental systems encouraged subjects to interact with the results of their search. They show many representations of the top-ranked documents directly to the subject at the results interface. These interfaces aim to facilitate the swift resolution of information needs but since they are novel, depend on their usability. For this reason, the training strategy (described in Chapter Nine, Section 9.10) was important, as was subject reaction to the systems. In this section results are presented on the relevance paths and information displayed at the search interface.

11.4.1.1 Relevance Paths

Subject interaction with relevance paths was automatically logged by the experimental systems. In this section I present the results of this log data analysis. Table 11.8 shows the most common path taken, the average number of steps followed, the average number of complete and partial paths and the average number of occasions where a subject went straight to a document from the first representation they visited. All averages are for each group of subjects over all search tasks. A complete path involved a subject visiting all five document representations and *then* the document itself. In partial paths, subjects visit only some document representations and do not have to visit the source document. Analysis of this sort can reveal how subjects actually used a search system rather than their perceptions of its use.

Table 11.8

Use of relevance paths (range 1-5, lower = better).

Factor	Inexperienced			Experienced		
	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
Most common path	TRS ↓ Title	Title ↓ Summary ↓ Summary Sentence		TRS ↓ Title ↓ Summary	Title ↓ Summary ↓ Summary Sentence ↓ Summary Sentence in Context	
Average steps	2.32	3.08	3.10	3.63	4.38	4.41
Average complete (partial) paths	5.20 (13.30)	7.35 (11.33)	7.54 (11.90)	7.32 (17.54)	11.64 (15.10)	11.85 (15.32)
Straight to document	6.61	6.35	6.48	9.62	9.30	9.76

n(inexperienced) = 24, *n*(experienced) = 24

Experienced subjects interacted more with the results of their search. Their paths were generally longer and they also followed more complete and partial relevance paths. They also went directly to more documents than the inexperienced subjects. These differences between groups were significant with a Mann-Whitney Test ($U(24) = 417$, $\alpha = .0167$, $p = .004$) for each pair-wise comparison (e.g., average steps (*inexperienced*/S_{Check}) *versus* average steps (*experienced*/S_{Check})). The option to directly indicate which items are relevant had an obvious effect on the interaction of experimental subjects. In the Checkbox system both subject groups interacted with shorter relevant paths than the Recommendation and Automatic systems. All users of the Checkbox system followed less complete and more partial paths than the other systems (Friedman Rank Sum Test, $\chi^2(2) = 12.43$, $p = .002$). This could be because subjects were trying to identify which representations were relevant rather than engaging themselves fully in their search.

There were only minor differences in the use of relevance paths for different task categories. I posit that the 15 minute task time was insufficient for real differences in subject search behaviour to emerge. Those studies that have found different search behaviours for different stages in the information seeking process (e.g., Kuhlthau, 1991) have been longitudinal and have monitored search behaviours over a period of weeks and months. While subject perceptions of the tasks differed, there was insufficient evidence from their interaction to suggest they interacted differently.

11.4.1.2 Content

To test the value of the interfaces to the experimental systems, subjects were asked about how the information was presented at the results interface. A set of four semantic differentials were used to elicit subject opinion: ‘helpful’/‘unhelpful’, ‘useful’/‘not useful’, ‘effective’/‘ineffective’, ‘distracting’/‘not distracting’. This was an important question, if subjects did not perceive direct benefit from the interfaces it may have adversely affected how they used them. The average responses for the four semantic differentials are shown in Table 11.9.

Table 11.9
Subject perceptions of information presented at the search interface
(range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
helpful	2.07	**1.96**	2.11	2.17	**2.14**	2.17	2.17	**2.05**	2.14
useful	2.29	2.29	**2.28**	2.18	**2.12**	2.08	2.33	2.20	**2.18**
effective	2.23	2.13	**2.10**	2.34	**2.26**	2.29	2.29	**2.19**	2.20
not distracting	2.38	2.21	**2.00**	2.28	2.18	**2.17**	2.28	2.19	**2.08**
all	2.25	2.15	2.13	2.24	2.18	2.18	2.27	2.16	2.15

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

The experimental systems presented information on the interface in the same way. Friedman Rank Sum Tests were applied within each subject group to test for statistical differences between the experimental systems and to see if components that varied between systems affected subject perceptions of the content shown. These tests revealed no significant differences in the value of the content presented between any of the experimental systems (all $\chi^2(2) \leq 2.93$, $\alpha = .0125$, all $p \geq .231$). Variations in interface provision for creating queries and making new search decisions therefore did not effect subject perceptions of how useful the content shown to them was. There were no significant differences between subject groups (Mann-Whitney Test, all $U(24) \leq 338$, all $p \geq .15$) and no significant interaction effects between search experience and systems ($\chi^2(2) = .77$, p = .68). In the next section the interface techniques used to reformulate the query are evaluated.

11.4.2 Term Selection

At any point in the search the experimental systems allowed the formulation of new query statements. When prompted, the Checkbox system presented the original query and the best non-query terms in a text box and allowed the subject to retain those terms added, add their own terms or remove terms to formulate the new query. The Recommendation system presents a list of recommended terms and allows the subject to add the best terms from this list to the query. The Automatic system generates a new non-editable query, but does allow the subject to create their own query for re-searching the Web. Subjects were asked to indicate on a Likert scale how comfortable they were with each query formulation method. The average responses are shown in Table 11.10.

Table 11.10

Subject perceptions of term selection methods (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
2.79	**2.13**	2.96	2.63	**1.96**	2.88	2.71	**2.04**	2.92

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

A Friedman Rank Sum Test was applied to the values in each group and the results indicated statistically significant differences in all groups (all $\chi^2(2) \geq 17.03$, all $p < .001$). Dunn's *post hoc* tests were applied to the data and revealed (in all three groups) significant differences between the Recommendation system and the other systems (all $Z = 3.12$, all $p < .001$). The differences between the Checkbox and Automatic systems were not significant in any groups (all $Z \leq 1.16$, all $p \geq .123$). In Chapter Four, the *TRSFeedback* study showed that relevance indications communicated implicitly could be a substitute for their explicit counterpart. This finding suggests in certain circumstances term selection components in such systems may also in some way be substitutable, and the case of the Recommendation system, perform better. There were no significant differences between subject groups (Mann-Whitney Test, all $U(24) = 353$, all p = .09) and no significant interaction effects between search experience and systems ($\chi^2(2) = 1.06$, p = .59).

The Likert scale analysed in Table 11.10 asks subjects to make a value judgement on the interface technique used to create the new query. Subjects appeared to like the presentation of the terms in a list separated from the query, allowing them to choose which terms were relevant and move these terms into the query. In the Checkbox system the new terms were included in the query box meaning the subject had to remove those that were not relevant. Also, the Checkbox system required subjects to explicitly request support with query formulation, something they forgot about or appeared unwilling to do. Experimental subjects generally did not like these additional burdens. In the next section I present and analyse findings on the interface support mechanisms for retrieval strategy selection.

11.4.3 Retrieval Strategy Selection

The experimental systems implemented retrieval strategies to gather a new set of documents or restructure the information already retrieved. The Automatic system follows strategies on behalf of subjects, the Recommendation system recommends them and the Checkbox system relies on the subject to choose them. In a similar way to the previous section, subjects were asked to indicate on a Likert scale how comfortable they were with the method used to select retrieval strategies in the experimental systems. Subjects' average response for each system, from each subject group, in shown in Table 11.11.

Table 11.11
Subject perceptions of retrieval strategy selection methods (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
2.23	**2.04**	2.92	2.21	**1.94**	2.63	2.22	**1.99**	2.78

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

A Friedman Rank Sum Test was applied to the values in each group and the results indicated the presence of effects in all groups (all $\chi^2(2) \geq 14.26$, all $p < .001$). Dunn's *post hoc* tests were applied to the data and revealed (in all groups) significant differences between all systems and all other systems (all $p \leq .001$). There were no significant differences between subject groups (Mann-Whitney Test, $U(24) = 350$, $p = .10$) and no significant interaction effects between search experience and systems ($\chi^2(2) =1.94$, $p = .38$). Subjects preferred the Recommendation and Checkbox systems since they had final control over how the revised query was used. The Recommendation system was preferred since as well as giving searchers control, it also made recommendations about which strategy should be followed; subjects could ignore or accept the recommendation. Later in this chapter I use interaction logs to analyse how many of the recommended actions were accepted. The Automatic system was not liked because it removed this control and intruded on subjects' search. The option to reverse all search decisions it made did not compensate subjects for the additional burden of having to do so.

The experimental systems used different methods to gather relevance information. Some gather assessments unobtrusively from subject interaction and others more directly. In the next section I analyse the results obtained when subjects were asked about the provision of relevance information in each of the three experimental systems.

11.4.4 Relevance Assessment

The experimental systems differ in how subjects could communicate which information presented at the interface was relevant. The Checkbox system presents checkboxes next to each representation and allows subjects to explicitly mark relevant items. The Recommendation and Automatic systems use implicit assessments of relevance, generated during subject interaction with the system. Subjects were asked about how they told the system which items (e.g., titles, summaries, Top-Ranking Sentences) were relevant. Unlike traditional RF systems, subjects were not able to mark whole documents as relevant; instead they assessed representations of documents. This may allow them to make more accurate relevance assessments.

They were asked to complete two semantic differentials about:

1. the *effectiveness* of the assessment method i.e., *How you conveyed relevance to the system was*: ‘easy’/‘difficult’, ‘effective’/‘ineffective’, ‘useful’/‘not useful’.
2. how subjects *felt* about the assessment method i.e., *How you conveyed relevance to the system made you feel:* ‘comfortable’/‘uncomfortable’, ‘in control’/‘not in control’.

The average obtained differential values are shown in Table 11.12 for inexperienced subjects, experienced subjects and all subjects, regardless of search experience. The value corresponding to the differential ‘all’ represents the mean of differentials one and two for a particular experimental system.

Table 11.12
Subject perceptions of relevance assessment methods (range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
easy	2.46	1.88	**1.79**	2.46	2.00	**1.96**	2.46	1.94	**1.88**
effective	2.75	**1.96**	2.67	2.63	**2.18**	2.67	2.69	**2.07**	2.67
useful	2.50	**2.13**	2.42	2.46	**2.14**	2.40	2.48	**2.12**	2.41
all (*diff. 1*)	2.57	1.99	2.29	2.52	2.11	2.34	2.55	2.05	2.32
comfortable	2.46	**1.88**	2.21	**2.14**	2.21	2.26	2.30	**2.05**	2.23
in control	**1.96**	2.25	3.21	**1.98**	2.13	3.14	**1.97**	2.19	3.13
all (*diff. 2*)	2.21	2.06	2.71	2.06	2.17	2.70	2.13	2.12	2.68

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Friedman Rank Sum Tests were applied within each subject group (differential 1: $\alpha = .0167$, differential 2: $\alpha = .0250$). The results of this analysis suggested significant differences in all semantic differentials and all subject groups (all $\chi^2(2) \geq 10.60$, all $p \leq .005$) except the 'comfortable'/*experienced* comparisons ($\chi^2(2) = 4.21$, $p = .122$). Experienced subjects appear equally comfortable with the relevance assessments in all systems.[39] Their search experience may allow them to adapt between interface technologies more easily. Dunn's *post hoc* tests were run on all differentials revealing significant differences for all comparisons (all $Z \geq 2.26$, all $p \leq .012$). These differences suggest that subjects found the implicit methods easy and useful in their search. In the Checkbox system subjects could decide which document representations were marked as relevant. Subjects felt more in control when given the additional responsibility for communicating relevance but, for inexperienced subjects, not necessarily more comfortable. Inexperienced subjects found the explicit communication of relevance difficult. Subjects with less search experience may find it problematic to adapt to new techniques for controlling their search.

The Recommendation and Automatic systems used implicit feedback techniques to estimate which information was relevant. These systems made inferences about information needs directly from search behaviour. The systems assume that when searching for information a user will try to maximise their rate of gain of relevant information. This assumption is at the centre of *information foraging theory* (Pirolli and Card, 1995), and assumes: (i) that the examination of documents and related information is driven by information needs, and; (ii) that searchers will try to maximise their rate of gain of relevant information whilst minimising the amount of irrelevant information. To test whether information needs drove interaction in the experimental systems, subjects were asked to indicate on a Likert scale the extent to which they agreed with the statement: *As I searched, I tried to only view information related to the search task.* The average Likert scale responses are presented in Table 11.13.

Table 11.13
Subjects tried to view relevant information (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
1.71	**1.67**	1.78	1.71	**1.50**	1.62	1.71	**1.59**	1.70

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

For the Recommendation and Automatic systems, these findings were important since they operate under the assumption that subjects will look try to view relevant information as they

[39] There was an interaction effect between search experience and the experimental systems for the 'comfortable' differential ($\chi^2(2) = 7.38$, $p = .025$).

search. Friedman Rank Sum Tests were applied and suggested no significant differences between systems for inexperienced subjects ($\chi^2(2) = 2.69$, $p = .261$) but there were for experienced subjects ($\chi^2(2) = 6.95$, $p = .031$). Dunn's *post hoc* tests revealed differences between the systems that gathered relevance information implicitly and the Checkbox system. Experienced subjects may have been able to infer how the Recommendation and Automatic systems choose additional terms (i.e., through the document representations viewed). There were no significant differences between subject groups (Mann-Whitney Test, $U(24) = 356$, $p = .08$) and no significant interaction effects between search experience and systems ($\chi^2(2) = .58$, $p = .75$). In the post-experiment 'Exit' questionnaire a number of experienced subjects explained that they had tried to be selective with the information they viewed since they assumed this must be how the systems that use implicit feedback gathered their evidence. That is, experimental subjects' perceptions of system operation influenced their interaction.

To assume that all the information a subject expresses an interest is in relevant may be too coarse grained since subjects can also interact with non-relevant information. To investigate the validity of this claim, interaction log data was used to calculate the proportion of all possible representations in the top 30 retrieved documents used to construct representations that were relevant (i.e., the search precision). In the Checkbox system this is the proportion of all possible representations that were marked relevant by the subject. Precision is computed in the Recommendation and Automatic systems based on the proportion of all possible representations that the subject expresses an interest in. The average number of document representations created or extracted from the top 30 documents was 320.65. There are a maximum of 14 representations per document; four Top-Ranking Sentences, one title, one summary, four summary sentences and four summary sentences in document context. However, since representations are created based on document content there is a chance that the documents may contain insufficient text to extract four sentences or may take too long to download. The precision values are shown in Table 11.14 and in Figure 11.3. For the Checkbox system the potential precision value is also given (in brackets) if implicit assessments had been used.

Table 11.14

Average search precision (values are percentages).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
1.25 (20.96)	21.65	21.36	2.76 (17.05)	17.17	16.52	2.01 (19.01)	19.41	18.94

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

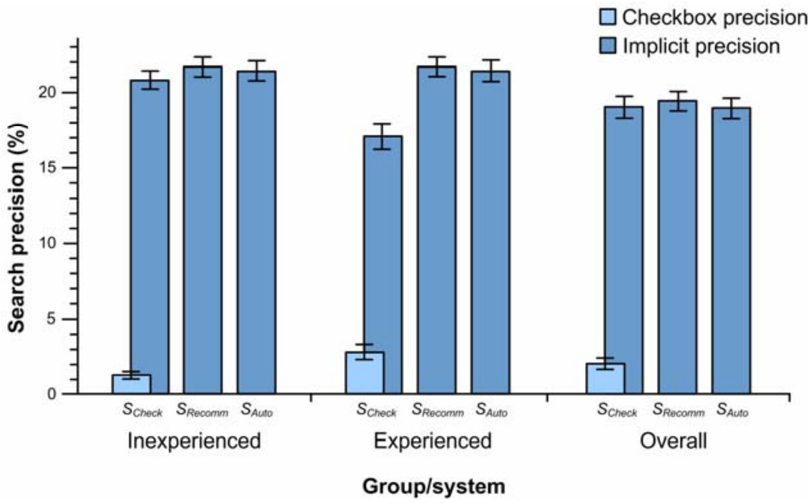


Figure 11.2. Search precision across system type and subject group (+/– *SE*).

The average search precision values shown in Table 11.14 suggest large differences in the number of items marked relevant in the Checkbox system and those inferred relevant in the Recommendation or Automatic systems. Subject criteria for marking a representation was generally very strict. During the experiment subjects suggested that an item had to be definitely relevant before they marked it. The Checkbox precision values differ significantly from those of the Recommendation and Automatic systems for both subject groups and overall (Wilcoxon Signed-Rank Test, all $T(24) \geq 229$, all $p \leq .012$). The precision values for the Recommendation and Automatic are very similar and do not differ significantly between subject groups (Mann-Whitney Test, $U(24) = 351$, $p = .097$). From these results it is obvious that experienced subjects check more items yet look at fewer. This could be because they are interacting more efficiently or assessing the relevance of items more carefully.

The highest precision value in Table 11.14 is still less than one quarter of the possible representations in the top-ranked document set. The probabilistic framework tries to estimate subject interests based on terms extracted from these representations. The experienced subjects expressed an interest in less document representations than the inexperienced subjects. These differences were not significant with Mann-Whitney Test for both the Recommendation ($U(24) = 356$, $p = .08$) and Automatic systems ($U(24) = 365$, $p = .06$). Nonetheless, this partially supports the earlier claim that experienced subjects used the systems with implicit feedback more cautiously.

Subjects provided additional informal comments on the relevance assessment process during and after the experiments. From subject comments, three factors emerged as important when indicating which results were relevant: the *method* used to communicate, the *value* of the communication and the *criteria* used during the communication. The *method* describes how relevance indications were elicited at the interface and subjects typically forgot to provide these indications. The *value* describes the perceived benefit of conveying indications and subjects generally felt the process was not worth their effort. Finally, the *criteria* employed during the communication were typically strict (i.e., results had to be completely relevant) and subjects rarely found results they regarded as relevant. How these factors are addressed is a challenge for developers of search systems that allow subjects to make relevance indications. Subjects preferred implicit relevance assessments over explicit assessments. This is beneficial for the searcher as they no longer have to be burdened with the responsibility of providing relevance assessments and for the term selection models, who receive more evidence from which to make their decisions.

When the Recommendation and Automatic systems chose terms and made search decisions they notified the searcher by displaying messages and changing the state of interface components (e.g., colour, rank order). In the next section subject perceptions of these notifications are presented and analysed.

11.4.5 Notification

The Recommendation and Automatic systems recommended/chose new search decisions for the subject as they searched. They notified the subject through a message at the interface and by placing an 'idea bulb' next to the mouse cursor. In the 'Search' questionnaire, issued after tasks had been attempted on these two systems, subjects were asked to complete semantic differentials eliciting their opinion about these notification methods. The differentials asked about:

1. the *communication* of search decisions i.e., *The system communicated its action in a way that was*: 'unobtrusive'/'obtrusive', 'informative'/'uninformative', 'timely'/ 'untimely'.
2. the '*idea bulb*' i.e., *The appearance of the* '*idea bulb*' *when the system chose/recommended an action was:* 'not disruptive'/'disruptive', 'useful'/'not useful'.

The average differential responses for inexperienced subjects, experienced subjects and all subjects, regardless of search experience are shown below in Table 11.15.

Table 11.15

Subject perceptions of system notification methods (range 1-5, lower = better).

Differential	Inexperienced		Experienced		Overall	
	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
unobtrusive	**1.96**	2.42	**1.58**	1.67	**1.77**	2.04
informative	**2.21**	2.54	**1.92**	1.96	**2.06**	2.25
timely	**2.38**	2.58	**2.38**	2.88	**2.38**	2.73
all (*diff. 1*)	2.18	2.51	1.96	2.17	2.07	2.34
not disruptive	**1.71**	1.67	**1.42**	1.71	**1.56**	1.69
useful	**1.71**	2.00	**1.67**	2.00	**1.69**	2.00
all (*diff. 2*)	1.71	1.83	1.54	1.85	1.63	1.84

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Wilcoxon Signed-Rank Tests were applied within-subject groups (differential 1: $\alpha = .0167$, differential 2: $\alpha = .0250$). The results of this analysis showed that there were significant differences between systems for all differentials (all $T(24) \geq 227$, all $p \leq .014$). These results suggest that although subjects preferred the Recommendation system's notifications, how the Automatic system communicated its decisions were also effective. There were no significant interaction effects between search experience and the experimental systems used ($\chi^2(1) = 0.18$, $p = .67$).

At the end of the experiment subjects were asked to rank the experimental systems in order of preference. In the next section I analyse subject responses.

11.4.6 System Preference

Subjects used each of the three systems and were asked to rank them in their order of preference. No instructions were given on what factors to use when making their decision, but subjects were asked to explain their ordering. In Table 11.16 I present the rank order of the systems for each subject group and within this group for the different task types.

Table 11.16

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

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
2.00	**1.45**	2.46	2.25	**1.29**	2.46	2.13	**1.42**	2.46

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

A Kruskal-Wallis test was applied to the rankings in each subject group, and to both groups combined. The results presented in Table 11.16 showed significant differences in the rankings assigned by inexperienced subjects ($\chi^2(2) = 4.61$, $p = .010$), experienced subjects ($\chi^2(2) = 14.03$, $p < .001$) and all subjects ($\chi^2(2) = 16.22$, $p < .001$). A Dunn's *post hoc* test was used to perform multiple comparisons within each subject group. There was a significant difference in the inexperienced group between the Automatic and Recommendation systems ($Z = 2.23$, $\alpha = .0167$, $p = .013$). For experienced subjects and across all subjects there are significant differences in the ranks assigned to the Recommendation system and the other two experimental systems (all $Z \geq 2.65$, all $p \leq .004$). The Recommendation system is the preferred search system for both subject groups and overall across both subject groups.

Table 11.17
Rank order of systems per subject group and task category (range 1-3, lower = better).

System	Inexperienced			Experienced			Overall		
	Pre-focus	Focus form.	Post-focus	Pre-focus	Focus form.	Post-focus	Pre-focus	Focus Form.	Post-focus
Checkbox	2.45	2.01	**1.54**	2.63	2.58	**1.55**	2.54	2.30	**1.55**
Recommendation	**1.05**	1.45	1.85	**1.05**	1.35	1.48	**1.05**	1.40	1.67
Automatic	**1.95**	2.66	2.76	**1.95**	2.66	2.76	**1.95**	2.66	2.76

n(inexperienced) = 8, *n*(experienced) = 8, *n*(overall) = 16

The effect of the different task categories on the ranking was also analysed. All within-group differences were significant (i.e., horizontally within group) (Friedman Rank Sum Test, $\chi^2(2) \geq 10.60$, $\alpha = .0167$, $p \leq .005$). There were no interaction effects between search experience and task categorisation (Friedman Rank Sum Test, $\chi^2(2) \geq 1.06$, $p \geq .59$). Each cell in the bottom three rows of Table 11.17 represents the average rank assigned by the subjects that attempted a task from that task category on that system. The results appear to indicate an association between task complexity and system preference, with systems that remove aspects of searcher control (i.e., Recommendation and Automatic system) being preferred for more complex search tasks and those that give searchers more control being preferred for less complex tasks (i.e., Checkbox system). However, since the number of trials in each cell is relatively small one must be conservative in any conclusions drawn from these results.

The reasons subjects gave for their rankings were also analysed. In a similar way as search tasks in Section 11.3.3, the reasons given by subjects are shown in Table 11.18.

Table 11.18
Subject comments on experimental systems
(numbers in brackets reflect the concept/statement frequency).

Checkbox	Recommendation	Automatic
1. "too much control"	1. "in control" (3)	1. "simple" (5)
2. "complex – better if user knows what they want"	2. "gives help, not over the user"	2. "too little control" (5)
3. "clunky"	3. "easy to operate...intuitive"	3. "not comfortable with results"
4. "too much hassle"	4. "non-obtrusive...no hassle"	4. "too objective"
5. "slow"	5. "good balance" (2)	5. "made user feel passive"
6. "too many choices" (4)	6. "didn't like choosing terms"	6. "a lot quicker"
7. "too many checkboxes"	7. "felt good!"	7. "least flexible system" (2)
8. "checking boxes is tiresome" (2)	8. "perfect blend"	8. "frustrating" (2)
9. "simple to use...felt in control"	9. "felt inclined to try [new words]"	9. "little indication of what system was doing" (3)
10. "a lot of effort" (2)	10. "simple to use, actions slightly unpredictable"	10. "not useful at all"
11. "concentrated on looking for information than checking boxes"	11. "powerful search options"	11. "no way of asking for a recommendation"
12. "forget to check boxes"	12. "didn't interfere"	
13. "added another dimension to search that could become frustrating"	13. "felt personal, as if it was understanding me"	
14. "a bit tedious"	14. "gain new insights and words"	

n = 48

Table 11.18 presents a general overview of comments provided by the experimental subjects. The Recommendation system receives mainly positive comments and the Checkbox and Automatic systems mainly negative. The Checkbox system offers too many options, increased the burden on the subject and interfered with the process of finding information. The consensus among subjects is that the Checkbox and Automatic systems do have good qualities: for the Checkbox system it is the control over which results are marked relevant, for the Automatic system it is the simplicity and control of the search. [40] However, despite these qualities subjects prefer the Recommendation system to the other systems.

In this section results have been presented and analysed for the first hypothesis. The results have shown that subjects preferred the experimental system that recommended terms and retrieval strategies. Subjects found the Checkbox system a hindrance in their search, that it presented them with too many choices and that it added an additional component to the search process that could become frustrating. The Automatic and Recommendation systems

[40] One subject remarked after a successful search on the Automatic system "maybe the system was better off being in control!".

provided a mechanism through which relevance information could be conveyed that was found to be straightforward and did not disrupt subjects' search patterns. Subjects were asked informally about the activity of creating queries in each of the three experimental systems; they preferred being able to select the terms used in the creation of their query. They did not like the Automatic system which did not let them refine their query for certain system operations. The selection of query words is an activity for which subjects want support from the system in proposing additional keywords. They suggested that this could be helpful where they may not be able to create good queries. Subjects viewed the creation of a new query as an important activity that they would rather control.

Subjects were also asked about selecting search strategies. The Automatic system removed all subject responsibility for selecting strategies. In a similar way to how they felt for query creation, subjects wished to retain control over the strategies employed, but responded well to recommendations made by the systems. For strategies that restructured retrieved information rather than recreating it, subjects were more willing to delegate control to the search system. That is, the amount of control subjects wished to retain was based on the predicted impact of the strategy.

In the next section I continue my analysis and present the results used to test the second experimental hypothesis.

11.5 Hypothesis 2: Information Need Detection

The second experimental hypothesis was that: *subjects found the terms chosen by the probabilistic implicit feedback framework valuable and worthwhile*. To test this hypothesis I analysed the *value* (can be helpful during a search) and *worth* (is correct and accurate) of the terms chosen by the framework. All experimental systems chose terms using the term selection model based on Jeffrey's rule of conditioning described in Chapter Seven. The results presented in this section therefore contribute to a test of the model rather than the experimental systems. Results are presented on a per system basis to test whether the interface support affected subject perceptions of the terms selected.

In Chapter Eight the retrieval effectiveness and the rate of 'learning' of the Jeffrey's term selection model was established with searcher simulations, independent of human subjects. The 'Search' questionnaire contained a section devoted to testing this hypothesis. Subjects were asked to answer a variety of semantic differentials, Likert scales and other question types. These data collection methods were used to gauge the effectiveness of the term selection model from the subjects' perspective.

11.5.1 Perceptions and Actions

Subjects were asked to complete two semantic differentials on whether the terms chosen by the system were 'relevant'/'irrelevant' and 'useful'/'not useful'. The average differential values are presented in Table 11.19 grouped by subject group.

Table 11.19
Subject perceptions of terms chosen/recommended by the experimental systems
(range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
relevant	2.58	**2.25**	2.63	2.33	**2.04**	2.38	2.46	**2.15**	2.50
useful	2.88	**2.38**	2.88	2.33	**2.17**	2.29	2.61	**2.27**	2.58
all	2.73	2.32	2.78	2.33	2.10	2.33	2.53	2.21	2.54

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Friedman Rank Sum Tests were applied to each differential for each group. The result suggested the existence of significant differences (all $\chi^2(2) \geq 7.54$, $\alpha = .025$, all $p \leq .023$). No Dunn's *post hoc* tests revealed significant differences in all subject groups between the Recommendation system and other experimental systems (all $Z \geq 2.17$, all $p \leq .015$). This suggests that the subjects perceive the terms recommended by the Recommendation system to be more relevant and useful. Although the same term selection model is used to choose terms, the data in Table 11.8 shows that the subjects interact more with the Recommendation system, providing it with more evidence. This suggests that subjects did not notice a difference in the quality of the information retrieved between systems. The differences between subject groups were significant ($U(24) = 385$, $p = .023$) suggesting that experienced subjects responded more positively to the terms selected. There were no interaction effects between systems and search experience ($\chi^2(2) = 1.88$, $p = .39$).

To build effective query modification techniques and improve the model in future work, it is vital to not only establish which terms were relevant, but *why* they were relevant. The Checkbox and Recommendation systems offered additional terms to the subject. These terms were presented in such a way that they could be edited. I regarded the act of not removing a term (Checkbox system) and moving a term from the recommended list into the query (Recommendation system) as a sign of acceptance of that term. Subjects were asked to explain why they had accepted any of the terms recommended to them. The options available were that: 'they meant the same', 'related to words chosen already', 'could not find better words', 'represented new ideas', 'other'. Subjects were told they could select as many options as were appropriate. In Table 11.20 the reasons given by all subjects for accepting terms recommended to them are presented.

Table 11.20
Reasons for accepting terms (values are percentages).

Reason	Inexperienced		Experienced		Overall		All
	S_{Check}	S_{Recomm}	S_{Check}	S_{Recomm}	S_{Check}	S_{Recomm}	
Meant same	**16.32**	14.29	**14.65**	12.82	**14.98**	13.58	14.27
Related words	**45.95**	40.48	**43.90**	43.59	**44.87**	41.98	43.40
No better words	**13.51**	11.90	**12.20**	10.26	**12.82**	11.11	11.95
New ideas	23.98	**30.95**	28.23	**30.77**	26.70	**30.86**	15.72
Other	0.24	**2.38**	1.02	**2.56**	0.63	**2.47**	1.26

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

The removal of the third system meant that the analysis must be applied for a 2 × 2 factorial design. Wilcoxon Signed-Rank Tests were applied to test the significance of the data within each subject group. Most differences were not statistically significant at this level (most $T(24) \leq 185$, $\alpha = .0125$, $p \geq .16$). However, the results suggest that the Recommendation system provides more new ideas than the Checkbox system ($T(24) = 234$, $\alpha = .0125$, p = .008). The larger number of terms offered or other aspects of the interface support may explain these differences. There were no significant differences between subject groups ($U(24) = 319$, p = .26) or interaction effects between search experience and system ($\chi^2(1) = .26$, p = .61). From these findings, I can propose that the relatedness to current query terms and the novelty of the concepts they embody are two of the main reasons why subjects accept terms chosen by search systems on their behalf.

In all systems subjects could modify their query at any point in the search. This would involve them selecting additional query terms based on tacit knowledge and their current search experience. A good term selection model should suggest relevant terms and suggest terms that initiate ideas for other terms. In this investigation subjects were asked to describe where the additional terms *they entered* originated. They could select one from 'list of terms suggested by the system', 'retrieved set of documents and extracted information', 'a combination of the first two' and 'other'. If subjects chose 'other' they were asked to provide more details. Table 11.21 shows the origins of new terms entered by the subject. The values in the table are percentages and the sum of each column is 100%.

Table 11.21
Origin of additional terms (values are percentages).

Source	Inexperienced			Experienced			Overall			All
	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	
System terms	8.33	**20.84**	16.67	**29.17**	20.84	**29.17**	18.75	20.84	**22.92**	20.83
Documents and Extracted Information	20.84	**25.00**	16.67	29.17	**33.33**	16.67	25.00	**29.17**	16.67	23.62
Combination of the above	**50.00**	45.83	45.83	12.50	**33.33**	12.50	31.25	**39.57**	29.17	33.33
Other	**20.83**	8.33	**20.83**	29.16	12.50	**41.66**	25.00	10.42	**31.24**	22.22

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Most subjects appeared to choose additional terms based on a combination of the terms chosen by the system and the documents and extracted information. This is a worthwhile finding as it shows the terms generated by the model are not only useful to represent current information needs but to facilitate their development. Friedman Rank Sum Tests were conducted for each differential within each subject group. The results implied the existence of statistically significant differences in each group (all $\chi^2(2) \geq 9.92$, $\alpha = .0125$, all $p \leq .007$). The high percentage of new ideas from 'other' sources (the percentages shown in the last row of Table 11.21) came from a combination of the simulated work task situation and the subject's tacit knowledge. The differences between the subject groups is significant for all differentials (all $U(24) \geq 392$, $\alpha = .0125$, all $p \leq .016$) . There is also evidence of interaction effects between the level of search experience and the experimental systems for the 'combination of the above' and 'other' differentials ($\chi^2(2) \geq 5.80$, $\alpha = .0125$, all $p \leq .002$). This suggests that the level of search experience affects where subjects get their terms and that this source varies depending on the experimental system.

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

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

A good term selection model should select terms on behalf of the subject that approximate their information needs. To be used effectively subjects must trust the systems to select appropriate terms. Subjects were asked whether they trusted the system to choose terms on their behalf. They completed a Likert scale to indicate the extent they agreed with the statement: *I would trust the system to choose words for me*. A summary of responses is provided in Table 11.22.

Table 11.22
Trust system to choose terms (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
2.19	**2.03**	2.48	2.19	**1.65**	2.19	2.19	**1.84**	2.34

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Friedman Rank Sum Tests were conducted for each differential within each subject group. The results suggested the existence of statistically significant pairs (all $\chi^2(2) \geq 11.24$, all $p \leq .001$). Dunn's *post hoc* tests revealed that there were significant differences in all inexperienced comparisons and for the experienced and overall subject groups, the Recommendation/Automatic (*experienced*: $Z = 2.03$, $p = .021$; *overall*: $Z = 2.00$, $p = .023$) and Recommendation/Checkbox (*experienced*: $Z = 2.05$, $p = .020$; *overall*: $Z = 1.90$, $p = .029$). Subjects appear to trust systems that give them control over query modification more than those without this facility.

Subjects were encouraged to provide comments on the terms suggested by all three systems. In general the feedback received was encouraging. Some subjects complained that certain terms and their plural appeared in the query suggested by the system (e.g., 'mite', 'mites'), this was unhelpful. On the other hand, one of the search tasks involved looking for art galleries in Rome. Since some of the retrieved pages were in Italian the system would occasionally suggest Italian words (e.g., 'galleria', 'museo') that were regarded by subjects as useful for their search. The system is therefore suggesting terms that the searcher may be incapable of selecting. In general subjects responded well to the terms chosen or recommended by the framework. The terms selected were helpful in either reinforcing current ideas or providing new ideas from which to advance their search. In the next section the interaction logs generated by each experimental system are analysed to provide further insight into how the framework was used in this experiment.

11.5.2 Query Construction

I this section I use interaction logs generated by each system to further investigate the creation of query statements. Since each experimental system supports different term selection strategies then different log data is available for each system. In this section the results from system logs are presented. The Automatic system does not allow the user the option of directly changing the new query. For this reason the logs analysed in this section are from the Checkbox and Recommendation systems.

Both systems use the probabilistic framework (Chapter Seven) for selecting query modification terms. The Checkbox system relies on the subject to mark items as relevant then suggests new query terms when instructed. The Recommendation system uses implicit feedback and recommends a list of terms to the subject.

11.5.2.1 Checkbox system

Unlike the other experimental systems, the Checkbox system awaits instruction from the subject before offering assistance. When requested, the system chooses the best six terms and appends them to the current query. The searcher then has the option to edit the query; adding or removing terms. I regard the removal of a term from those added by the system as a sign of dissatisfaction with the term (and its retention as a sign of satisfaction). Therefore, I use the proportion of terms added/removed from the original query as an indication of satisfaction/dissatisfaction with the term selection component of the probabilistic implicit feedback framework. Across all tasks on the Checkbox system an average of 2.15 of the six terms (35.83%) were rejected and 3.85 (64.17%) of terms were retained.

11.5.2.2 Recommendation system

The Recommendation system presents a list of recommended terms and allows subjects to choose terms from this list and add them to their query. This list contains 20 terms and it is unreasonable to expect subjects to add all 20 terms to their query. [41] It is also unreasonable to

[41] Due to limits with the underlying search system, a query used to re-search the Web cannot be any longer than 10 terms.

measure the proportion of the list that is selected as this is not comparable with the results in the previous section. Instead, I consider the quartiles of the list, and *where* in the list the terms reside. In a similar way to Efthimiadis (1993) it is possible to measure the proportion of offered terms added from the four quartiles of the list (top, top-middle, bottom-middle, bottom). Figure 11.4 shows how the top 20 terms were divided into four quartiles. The part of the list in the figure with a scrollbar represents the six terms shown to the searcher at any particular time. Subjects were not told that the terms in the list were ranked in descending order meaning they may not expect higher ranked terms to be more relevant. This allowed a more robust analysis of the list ordering, as if subjects chose more terms from near the top it would be because they thought they were useful, not that they assumed they should be.

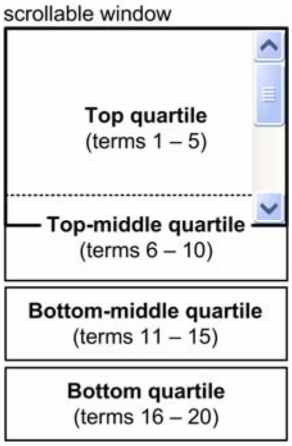


Figure 11.4. Four quartiles of the Recommendation system term list.

Since the terms are ranked by the framework, the location of terms in the list can give a clue about how well the term selection model operates. In Table 11.23 the proportion of terms chosen from each quartile in the list is shown for different subject groups and overall across all subject groups. The values in the table are percentages of the whole list.

Table 11.23
Proportion of terms chosen from list quartiles (Recommendation system only).

Quartile	Inexperienced	Experienced	Overall
Top	54.75	47.05	51.40
Top-middle	25.00	24.78	24.89
Bottom-middle	10.25	19.22	14.24
Bottom	10.00	8.95	9.47

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Mann-Whitney Tests were conducted for each quartile between each subject group (α = .0125). The results were significant for the top ($U(24)$ = 401, p = .001) and bottom-middle

(*U*(24) = 401, p = .001) quartiles, but not for the top-middle (*U*(24) = 321, p = .248) or bottom (*U*(24) = 341, p = .137). Overall subjects chose more than half of the recommended terms from the top five and over three-quarters (75.29%) from the top 10 terms (i.e., top and top-middle quartiles collectively). This implies that the subjects generally agreed with the ranking of terms by the term selection model.

To allow for the differences between the number of terms presented and more fully evaluate the recommended list of terms I ignore the scrollbar and only analyse terms with an initial rank position in the first six i.e., only terms that initially appear in the recommended list without the need to scroll. This meant that term selection methods in the Checkbox and Recommendation systems could be compared. Across all tasks and subject groups an average of 3.95 terms from the top six terms (65.85%) were added to the query. Analysis of these findings showed that although there was no significant difference between the number of terms added in the Recommendation and Checkbox systems (with a Wilcoxon Signed-Rank Test, *T*(24) = 198, p = .087). The way that additional terms are offered to subjects at the interface appears to only have a slight effect on the number of terms accepted.

11.5.2.3 Automatic system

Other than re-searching the Web, there was no mechanism for direct query refinement in the Automatic system. Subjects could modify and submit a new query to the system (i.e., re-search the Web), but received no support in choosing the terms to comprise this query. The queries submitted by subjects for the re-searching operation were typically smaller in this system (where the subject received no support) than in the other experimental systems which offered subjects assistance (2.53 terms *versus* 5.43 terms). The systems that implemented mechanisms for interactive query modification allowed subjects to build richer queries for generating new sets of search results.

In this section I have presented and analysed findings to test the second experimental hypothesis. The results have shown that the term selection model in the probabilistic framework chooses terms that are relevant and useful to subjects. The results also show that the nature of the interface support can affect subject perceptions of model effectiveness, including how much trust they place in it to choose terms on their behalf. In the next section I present and analyse results on the component used to estimate information need change that is used in the Recommendation and Automatic systems.

11.6 Hypothesis 3: Information Need Tracking

This section presents results related to the third experimental hypothesis: *subjects found the retrieval strategies chosen by the probabilistic implicit feedback framework valuable and worthwhile*. In the Recommendation and Automatic systems a component works in the background to suggest or choose new retrieval strategies during the search. These strategies are selected based on the extent of changes in the search system's formulation of information needs (i.e., changes in the list of candidate terms from which the system chooses query modification terms). To do this, the system uses the information need tracking component from the probabilistic implicit feedback framework described in Chapter Seven. As suggested in earlier chapters the framework can either re-search the Web or reorganise the information already retrieved. I test the effectiveness of this component using Likert scale and semantic differential responses, system logging (e.g., the proportion of system search decisions that are accepted by the subject) and informal subject comments.

11.6.1 Perceptions and Actions

In the 'Search' questionnaire, completed after each search task, subjects were asked to indicate on a five point Likert scale how often the retrieval strategy chosen by the framework reflected the changes in the information they were searching for. In the training session it was made clear to subjects that this change did not have to be a change in topic, it could simply be a refinement of their current search. They were asked to provide an assessment on a scale between 'never' and 'always'. The average scale responses are shown in Table 11.24.

Table 11.24
Subject perceptions on the appropriateness of retrieval strategy (range 1-5, lower = better).

Inexperienced		Experienced		All	
S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
2.54	2.58	**2.67**	2.71	**2.60**	2.65

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

The within and between group differences were not significant (*within*: Wilcoxon Signed-Rank Tests, all $T(24) \leq 182$, all p ≥ .180; *between*: Mann-Whitney Tests, all $U(24) \leq 358$, all p ≥ .08) and there were no interaction effects between search experience and experimental systems ($\chi^2(1) = 0.26$, p = .61). Since the mechanism for selecting retrieval strategies was the same between systems it was expected that that subject perceptions of the strategies would be similar. This was the case, but subjects again appeared slightly more positive about systems that gave them ultimate control over interface decisions.

To further test the information need tracking component, subjects were asked about the retrieval strategy chosen or recommended by the experimental system. A set of three semantic differentials were used to elicit subject opinion: 'useful'/'not useful', 'helpful'/'unhelpful', 'appropriate'/'inappropriate'. The strategy chosen by the system reflects changes in the system's estimation of the information need. The responses for the three differentials are shown in Table 11.25.

Table 11.25
Subject perceptions of retrieval strategies (range 1-5, lower = better).

	Inexperienced		Experienced		Overall	
Differential	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
useful	**2.38**	2.79	2.25	**2.21**	**2.31**	2.50
helpful	**2.54**	2.75	2.42	**2.21**	**2.48**	**2.48**
appropriate	**2.50**	2.92	**2.25**	**2.25**	**2.38**	2.58
all	2.47	2.82	2.31	2.22	2.39	2.52

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

The 'useful' and 'helpful' differentials in Table 11.25 measure the value of the strategy, i.e., how can the strategy assist subjects to search more effectively, and the 'appropriate' differential measures its worth, i.e., how well it performs. Wilcoxon Signed-Rank Tests were applied for each differential between systems. The tests revealed significant differences within the inexperienced subject group ($T(24) = 246$, $\alpha = .0167$, $p = .003$**)** but not the experienced group ($T(24) = 209$, $\alpha = .0167$, $p = .047$**)**. Inexperienced subjects found the retrieval strategy chosen by the Recommendation system significantly more 'useful' ($Z = 2.58$, $p = .005$), 'helpful' ($Z = 2.26$, $p = .012$) and 'appropriate' ($Z = 2.41$, $p = .008$) than the Automatic system. This was anomalous since the systems used the same underlying mechanisms to choose retrieval strategies. The only difference between the systems was in how the strategy was communicated. For inexperienced subjects, the method used to communicate the decision influenced subject perceptions about the value of the strategy. Experienced subjects seem more able to isolate the mechanism behind the strategy selection and no significant differences between the differentials were discovered for that group (all $Z \leq .74$, all $p \geq .23$). That is, experienced subjects were more able to analyse the value and worth of the information need tracking component independent of the way the decisions it made were communicated.

A good information need tracking component should choose retrieval strategies that approximate changes in the information needs of searchers and assist them in finding relevant information. To be used effectively, searchers must trust the systems to select appropriate retrieval strategies. Subjects were asked whether they trusted the system to choose retrieval strategies on their behalf. They completed a Likert scale to indicate the extent they agreed with the statement: *I would trust the system to choose an action* [42] *for me*. A summary of responses is provided in Table 11.26.

Table 11.26
Trust system to choose retrieval strategy (range 1-5, lower = better).

Inexperienced		Experienced		Overall	
S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
2.67	2.92	**2.67**	2.67	**2.67**	2.79

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Wilcoxon Signed-Rank Tests were applied within each subject group to compare systems and all subjects and systems compared to the mid-value of the Likert scale (i.e., 3). The results showed no significant within-group differences (all $T(24) \leq 160$, all $p \geq .390$), significant differences from the mid-value ($T(24) = 229$, $p = .012$) and no interaction effects between search experience and experimental systems ($\chi^2(1) = 0.15$, $p = .70$). Subjects reacted positively to the search strategies proposed by the system. Inexperienced subjects appeared to trust systems that gave them control over how the new query was used; for experienced subjects there was no difference.

In this section I have presented an analysis of subject perceptions of the retrieval strategy selection component. In the next section, I use system log data to analyse how subjects actually selected retrieval strategies. These logs, created as subjects searched, provide evidence to allow a deeper analysis of subject search activities.

11.6.2 Retrieval Strategy Selection

The Recommendation and Automatic systems make search decisions on subjects' behalf, whereas the Checkbox system relies on subjects to make their own decisions. Subjects are given the option to reverse the search decisions the systems made. In Table 11.27 I give the proportion of each type of action that was reversed. This reversal is regarded as an indication of dissatisfaction with the outcome of followed strategy.

[42] The word 'action' is used in the questionnaires rather than 'retrieval strategy' or 'search decision'. It was felt that subjects could relate better to 'action'.

Table 11.27
Proportion of retrieval strategies accepted or reversed (values are percentages).

Subject Action	Inexperienced		Experienced		Overall	
	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
Accepted	72.43	**75.60**	64.67	**69.10**	68.55	**72.35**
Reversed	**27.57**	24.40	**35.33**	30.90	**31.45**	27.65

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

The differences between the systems within the subject groups are not significant (Wilcoxon Signed-Rank Test, all $T(24) \leq 156$, all $p \geq .431$) but it is between groups (Mann-Whitney Test, all $U(24) = 399$, all $p \leq .011$). Experienced subjects tended to accept a lower number of retrieval strategies chosen by the system than inexperienced subjects. These subjects may be more reticent about search systems making decisions of this nature on their behalf and feel able to make such decisions on their own.

I use a measure known as *strategy overlap* to determine how closely the decisions made by the information need tracking component concord with subject decisions. I measure the degree of strategy overlap using the Checkbox system and the Recommendation system. The methods used in each system are slightly different. In the Checkbox system the strategy selection component runs in the background, completely invisible to the subject and not involved directly in any strategy selection decisions. That is, whilst the component chooses retrieval strategies based on changes in its formulation of information needs, these strategies are never shown to the subject and never executed. At any point in time, the component holds that retrieval strategy that it regards as most appropriate. I measure the degree of strategy overlap based on how frequently subjects choose the same strategy as the system would choose. In the Recommendation system the overlap is a measure of how many strategies followed by the subject that were also the system's recommendation at that time. This is different from the results reported in Table 11.27, since for this analysis I do not consider whether the strategy was eventually reversed or accepted. This is given as a percentage and is presented in Table 11.28 for inexperienced subjects, experienced subjects and across all subject groups.

Table 11.28
Proportion retrieval strategy overlap between system and subject (values are percentages).

System	Inexperienced	Experienced	Overall
Checkbox	61.60	57.85	59.73
Recommendation	74.66	59.32	66.99

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

On approximately 60% of occasions the framework implemented in the Checkbox system predicted the strategy executed by the subject. The differences are not significant between subject groups with a Mann-Whitney Test (*U*(24) = 345, p = .120). This is a reasonable result since the evidence gathered to predict the changes that result in the strategy are based on a small amount of evidence explicitly provided by the subject through their interaction. The strategy overlap for the Recommendation system is higher than the Checkbox system. There are at least two reasons for this: (i) since it gathers relevance assessments implicitly the system has more relevance information from which to make its decisions, and (ii) the presentation of the recommendation at the interface may have unduly influenced subjects into selecting it. The inexperienced subjects follow significantly more of the system's recommendations than the experienced subjects (Mann-Whitney Test, *U*(24) = 417, p = .004). They may require the additional support or be less cautious than the experienced subjects about accepting it. The Checkbox system may give an artificially low strategy overlap (because of the small amount of evidence) and the Recommendation system an artificially high value (because of the influence of presenting its decisions). Therefore, I conjecture that a 'true' strategy overlap value may well lie somewhere between these two extremes.

In this section the information need tracking component of the probabilistic implicit feedback framework has been tested. Subjects were asked to comment informally about the retrieval strategies. In a similar way to how they felt for query creation subjects wished to retain control over the strategies employed, but responded well to recommendations made by the system. For strategies that restructured retrieved information rather than recreating it, subjects were more willing to delegate control to the search system. That is, the amount of control subjects wished to retain was based on the predicted impact of the strategy. Subjects suggested that the component should be more sensitive to larger changes in information needs and that it reordered documents when their intuition would have been to re-search. Nonetheless, the component performed well and the results have demonstrated that the component makes search decisions that are appropriate and that subjects find useful.

11.7 Chapter Summary

In this chapter I have presented and analysed the findings of the user experiment. The experiment aimed to compare the effectiveness of three search interfaces that varied searcher control and responsibility over aspects of the search, and test the probabilistic implicit feedback framework presented in Chapter Seven. In Table 11.29 I summarise the results for each of the sub-hypotheses described in Chapter Nine.

Table 11.29
Evidence to support experimental hypotheses.

Hypothesis	Supported?	Evidence
Hypothesis 1. Interface Support		
Relevance Paths and Content (Hypothesis 1.1) Subjects find the information presented at the interface useful.	✓	Section 11.4.1
Term selection (Hypothesis 1.2) Subjects want control in formulating new queries.	✓	Section 11.4.2
Retrieval strategy selection (Hypothesis 1.3) Subjects want control in making search decisions.	✓	Section 11.4.3
Relevance assessment (Hypothesis 1.4) Subjects want the experimental system to infer relevance from their interaction.	✓	Section 11.4.4
Notification (Hypothesis 1.5) Subjects find system notifications helpful and unobtrusive.	✓	Section 11.4.5
Hypothesis 2. Information Need Detection		
Value (Hypothesis 2.1) Query modification terms chosen by the framework are relevant and useful.	✓	Section 11.5.1
Worth (Hypothesis 2.2) Query modification terms chosen by the framework approximate subject information needs.	✓	Section 11.5.1 Section 11.5.2
Hypothesis 3. Information Need Tracking		
Value (Hypothesis 3.1) The retrieval strategies chosen by the framework are beneficial.	✓	Section 11.6.1
Worth (Hypothesis 3.2) The retrieval strategies chosen by the framework approximate changes in the information needs of subjects.	✓	Section 11.6.2

The results have shown that subjects did not like having to mark items as relevant (as in the Checkbox system) or devolving control over query creation and retrieval strategy selection (as in the Automatic system). Subjects preferred to communicate relevance implicitly, and receive system support in creating queries and making new search decisions, but still retain ultimate control over these two activities. The Recommendation system offered them the facilities to do this. Hypothesis 1 was supported by these findings

In this chapter I also evaluated the probabilistic implicit feedback framework presented in Chapter Seven, to modify queries and select retrieval strategies. Subjects found the terms and strategies selected by the framework useful, relevant and appropriate in the context of their search. Hypotheses 2 and 3 were supported by these findings. In the next chapter I discuss the implications of the results obtained.

Chapter 12

Discussion

12.1 Introduction

In the previous chapter I presented and analysed the results of the user experiment. In this chapter these results are discussed in the context of this thesis and related literature; where appropriate, the findings are also compared to those of Pilot Test 1, described in Chapter Nine. In particular, I concentrate on results that relate to the three experimental hypotheses and other parts of this thesis. Each hypothesis is addressed in turn and this chapter concludes with a summary discussion of the implications of my findings.

Selecting worthwhile terms on behalf of searchers relies on an ability to predict their information needs to a very fine level of granularity. Traditional implicit and explicit relevance feedback approaches use sets of documents from which to extract terms for query modification (Salton and Buckley, 1990; Kelly and Teevan, 2003). This approach is coarse-grained since documents can contain a large number of erroneous terms (Allan, 1995). The approaches described in this thesis utilise interaction with novel content-rich search interfaces to modify the query statements and make search decisions.

Users of traditional search systems are typically responsible for all aspects of their interaction, from the selection of query terms to the assessment of the results obtained. This can be problematic as searchers typically receive no training in how to create queries, exhibit limited interaction with the results of their searches and do not examine results closely (Jansen *et al.*, 2000). The search interfaces presented in Parts II and IV use query-relevant document representations to facilitate access to potentially useful information and encourage searchers to closely examine search results. The findings in Part II showed that increased searcher interaction with retrieved information led to more effective searching. The interfaces in Part

II use the content of the most relevant documents in the retrieved set in an approach I call content-driven information seeking (CDIS).

IR systems that use implicit feedback make inferences about what information is relevant based on searcher interaction. They do not intrude on the searcher's primary line of activity (i.e., satisfying their information need). That is, the treatment by the system of the searcher's action as evidence of relevance is secondary to the main task, which is to respond to the searcher's instruction (Furnas, 2002).

RF systems typically have functionality for choosing query words, providing relevance information and making new search decisions. In this experiment I developed three experimental systems that tested these functions with subjects with different skill levels and search experience. This chapter begins with an initial discussion of the search process and search tasks attempted by subjects, then discusses interface support and the performance of the framework in detecting and tracking information needs.

12.2 Tasks and the Search Process

In this thesis I have described a number of user studies. Most of these studies have used simulated work task situations to facilitate interaction with the experimental systems. [43] These allow subjects to make personal assessments of what constitutes relevant information and allow search systems to be compared on the same underlying information need. In the studies described in Part II the subjects were not given a choice of tasks. This led to slight problems as some subjects were not interested in the task assigned to them. Borlund (2000b) recommended that in the construction of search tasks, experimenters should consider the involvement of subjects, the application of dynamic and individual information needs (real and simulated) and the use of multidimensional and dynamic relevance judgements. Subjects with an interest in the subject area of the task are more likely to become involved in the task and form an individual perspective of it. In Pilot Test 1 and in the experiment described thus far in Part IV I offered subjects a choice of search tasks that gave subjects more control over the tasks they attempted.

In Chapter Four I discussed the use of the top-ranking sentence based experimental interfaces in relation to the model of the Information Search Process (ISP) proposed by Kuhlthau (1991). This model assumes that there is a point of 'focus' (Kelly, 1963; Belkin, 1980; Kuhlthau, 1991) where uncertainty drops and searchers can better identify the topic of their

[43] With the exception of the *TRSFeedback* study in Chapter Four.

search. The findings from the user studies described in that chapter suggest that the systems support two of the six stages of the ISP: *exploration* (investigating information on general topic) and *collection* (gathering relevant or focused information). Those systems that used implicit feedback to reorder the Top-Ranking Sentences also displayed limited support for the *formulation* stage (formulating the search focus). However, since it is the system that is refining its formulation of the information need internally, the extent to which these systems support formulation (from the searcher's perspective) is limited. Through encouraging more interactivity in query creation, the systems presented in the experiment I have described in Part IV help searchers refine their query and improve support for the formulation stage of the ISP.

In this experiment, tasks were divided into three categories based on the actions common to each stage in the ISP. The tasks used in this evaluation simulate stages before the focus, as the focus is forming and after the focus has formed. The three task types created were assigned the names: *pre-focus*, *focus formation* and *post-focus*. The pre-focus tasks encouraged subjects to locate background information, the focus formation broadly relevant information and the post-focus broadly relevant or pertinent (focused) information.

Search tasks were created for each category using the approach described by Bell and Ruthven (2004) i.e., the task categories were varied in terms of complexity. The pre-focus task was assumed to simulate the state of an information need in the initial explorative stages of a search; encouraging browsing behaviour; this task was assumed to be highly complex. The focus-formation task simulated information needs as subjects began to understand what they were looking for and could then make decisions about what information was relevant; this task was assumed to be of moderate complexity. Finally, the post-focus task simulated a well-formed information need and encouraged focused information seeking; this task was assumed to have a low complexity. It is in the pre-focus stage where the information needs are least well-defined and most changeable.

I selected six search topics to approximate real information seeking scenarios. Subjects chose a task from each category without topic repetition to limit learning effects. This meant they choose the first task from six topics, the second from five topics (the first topic could not be attempted again) and the third from four (the first and second topics could not be attempted again). This methodology meant that the third topic selected could be a subject's third preference. The effect of this was negligible and was preferred to situations where topics were not removed (serious learning effects) [44] or subjects constructed their own search tasks (no comparability between systems). Unlike naturalistic studies (Beaulieu, 1997; Kelly, 2004) this investigation did not study natural search behaviours in operational settings. This experiment was a comparative evaluation and a deviation from a methodology where I could control many external factors could invalidate the experimental findings.

Subject comments in the 'Exit' questionnaire led me to conclude that they were able to identify differences between task categories. Subjects remarked that their search behaviour and task performance were affected by the nature of the task they attempted. It emerged from these comments that not only did subjects know that the task categories were different but that they also knew *how* the categories differed (i.e., in their complexity). There were no discernable differences in subject perceptions of the tasks between systems although subjects did find the pre-focus tasks more 'complex', 'unfamiliar' and 'unclear' than tasks from the other categories. Although there were only minor differences in subject interaction for each type of task, subject perceptions suggest that systems that gather relevance information using implicit feedback and make or recommend decisions (i.e., the Automatic or Recommendation systems) are most useful during the uncertain, formative stages of the information seeking process. Since these systems removed the burden of directly communicating relevance, subjects could focus on viewing and interpreting the documents and extracted information presented at the search interface. The Checkbox system was preferred in circumstances where subjects were more certain about the information they were searching for. When they become more aware of their information need, they felt more able to identify what information is relevant. In such situations they also *want* more control over system decisions and seem less reluctant to provide relevance feedback directly. This is a potentially significant finding, although since it does not form one of the hypotheses tested in this thesis a more complete analysis of the results are reserved for future work.

In a related study, Fowkes and Beaulieu (2000) suggested that the complexity of the search may be an indicator of when to use different query modification techniques. They found that for searches where the desired information is clearly defined and for which the searcher can retrieve relevant information they do not require as much control over the terms that comprise the query. Searches involving vague information needs or in cases where little relevant information is being retrieved benefit more from increased control over the query terms. However, in this experiment I demonstrate that the same may not be true for relevance assessments; subjects felt most comfortable directly communicating relevance to the

[44] With subjects being able to choose the same topic for all three tasks.

Checkbox system for less complex tasks. Since the direct communication of relevance information is dependent on an ability to identify what information is relevant, searchers may feel more comfortable doing this for less complex tasks. For more complex searches they may be unable to identify what information is relevant and may therefore rather rely on inferences made by the search system.

Subjects were asked about their perceptions of task success after they had attempted each search task. Task success was assessed from the subjects' perspective since this most closely reflects real life retrieval situations. Personal assessments on task completeness also fit better with the use of simulated work task situations, which require subjective judgements on what information is relevant. The findings of the experiment suggest that inexperienced subjects perceived higher levels of task success on the Recommendation system than any of the other two experimental systems. They commented that the way the system communicated its decisions (i.e., unobtrusively) meant they were not impeded in their search by a need to control the system. Subjects spent time on the Checkbox system assessing document representations for relevance, rather than searching for information and reversing or examining the effects of the Automatic system's decision. This reduced the amount of information they could examine during a search and for the more complex tasks lessened the likelihood of a successful search.

The results discussed in this section show that subjects noticed the differences in task complexity and that the experimental systems that were most proactive in offering searcher assistance were most useful for search tasks that encouraged explorative information seeking behaviour. The interface support mechanisms were the only differences between the experimental systems. In the next section I discuss findings about the first of the three research questions, which addresses the interface support issue.

12.3 Interface Support

In this section I discuss aspects of the interface support offered by the experimental systems. The three systems provided different mechanisms that varied how much control searchers had over aspects of their search. Many of the interface design decisions made for the experimental systems described in Chapter Ten arose from subject comments during Pilot Test 1. In that study two prototype experimental systems were tested: a manual baseline system and an experimental system that used implicit feedback. The manual baseline gave subjects control over the terms selected and retrieval strategy followed and the implicit feedback system automatically modified the query and used the new query to perform a new retrieval strategy. The implicit feedback system gave subjects no control over the process other than the option to reverse system decisions. This pilot test demonstrated that the heuristic-based implicit feedback framework (from Chapter Six) could approximate searcher interests and estimate changes in these interests (i.e., the framework worked well). However, subjects also suggested that they preferred systems that offered assistance in making search decisions, but gave them final control over the choices made. These comments were considered and influenced the development of the systems used in this experiment.

Systems with three different types of interface support were used to communicate the decisions on which terms and retrieval strategies were chosen by the underlying probabilistic framework. The way in which these decisions were communicated and the level of searcher control over them, was varied between systems. In the Checkbox and Recommendation system, new query terms were suggested as a recommendation and could be edited by the subject. In contrast, the Automatic system chose terms for the subject. Subjects generally preferred the interface support mechanisms provided by the Recommendation system.

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

A deeper understanding of what searchers want to control and what they are happy to delegate can assist in the development of more effective systems for interactive search. Techniques to facilitate the provision of relevance information, form new queries and use these queries were all tested in this experiment. The discussion of interface support is divided into three main parts: relevance indications, query creation and the selection of retrieval strategies. Each section begins with an italicised summary statement describing the main conclusion drawn.

12.3.1 Relevance Indications

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

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

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

In this experiment 'precision' was taken as a measure of search effectiveness and based on how much of the retrieved document set the subjects classed as relevant. To compute this measure, the Checkbox system used the proportion of potential representations [45] that were actually marked and the implicit feedback systems used the proportion of all representations that were classified as being relevant. The results suggested a large difference between how much information the implicit systems regarded as relevant and what the subject actually marked as being relevant. The relevance and usefulness of the terms generated from the implicit feedback systems was higher than that of the Checkbox systems, suggesting that more evidence, albeit less reliable than that provided by the searcher allowed better quality terms to be chosen by the implicit feedback framework. It also suggests that criteria subjects employed when assessing relevance was too strict and that better queries could have arisen from the selection of more representations that were perhaps not totally relevant. In the next section I discuss the interface techniques used to incorporate new query words.

12.3.2 Query Generation

Subjects preferred to retain control over query creation. The systems that allowed subjects to monitor and change the query were preferred over the Automatic system, which did not. They were willing to delegate the task of recommending potential keywords but not the task of adding these words. Subjects preferred control over the terms chosen by the system, even if this meant more work for them in moving terms of interest from the recommended term list to the query. This effort was seen to be both *unnecessary* (subjects were not forced to do it) and *worthwhile* (subjects perceived a benefit from it). The implicit nature of the evidence captured may make the search decisions of systems that use it unreliable and subjects may rather retain control to be sure of their correctness. Subjects engendered more trust in systems where they could verify the correctness of the words chosen prior to their submission. For more complex tasks they required more support in query formulation.

Subjects liked having terms suggested to them, but in a way that did not require them to delete irrelevant terms (as in the Checkbox system), only select relevant ones; subjects did not want to have to act to correct erroneous system decisions. Subjects were more willing to delegate responsibility for the creation of queries to systems that allow them to verify the correctness of system decisions. In a related study, Koenemann and Belkin (1996) tested search systems with different levels of visibility and interactivity in creating queries. In this experiment the Automatic system only allowed subjects to see the query created by the system; the Checkbox and Recommendation systems allow subjects to view *and* adjust the new query. In this

[45] All document representations in the top 30 documents that could be marked.

experiment, as in the work by Koenemann and Belkin, subjects preferred systems that gave them control over the new queries. That is, they want help in selecting query terms but want ultimately to decide which terms are used.

The Checkbox system chose terms for subjects based on the items they had marked as relevant. These items reflected their current information needs and the terms suggested by the system appeared to reflect these needs also. Subjects chose terms from those recommended in the Recommendation system because: (i) they represented new ideas, (ii) they meant the same as the query terms, and (iii) they were related to the query terms. The study by Koenemann and Belkin found that subjects tended to choose semantically related feedback terms. In this experiment I found that subjects use the query terms to give them ideas for what terms are appropriate or were related to the original terms in some way. For example, a search for 'worldwide petrol prices' could mean that the terms 'pipe', 'iraq' and 'dollar' are good feedback terms, but their semantic relationship to the original query is not immediately apparent.

All experimental systems tried to increase the length of subjects' query statements by expanding the original search query. Belkin *et al.* (2003) have demonstrated that experimental subjects can be more satisfied with search results if they submit longer queries to the search system. The use of a feedback system to choose terms on a searcher's behalf is only one way to create longer queries. Kalgren and Franzen (1997) demonstrated that a different style of query input box encouraged the submission of longer queries, a result verified by the Belkin *et al.* (2003). It is preferable to encourage searchers to better define their information needs. However, in circumstances where they may be unfamiliar with the topic of the search, they may be unable to produce longer queries (Kelly and Cool, 2002).

Traditional Web search systems are 'pull' oriented where it is the searcher's responsibility to locate relevant information. The systems I have described in this thesis operate on a 'push' paradigm and are adaptive, work to better describe information needs and consider changes in these needs, restructuring or recreating the information presented at the results interface. Once a new query has been generated it can be used to perform a *retrieval strategy*. In the next section I discuss the selection of such strategies.

12.3.3 Retrieval Strategy Selection

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

The Recommendation and Automatic systems dynamically update their internal representation of information need change and adopt the retrieval strategy to reflect the information need of the searcher, as estimated by the search system. Different search decisions had different levels of impact on a search. Reordering decisions restructured the already retrieved information at the interface, whereas re-searching decisions generated a new set of documents. The decisions increased in severity, from reordering Top-Ranking Sentences, to reordering documents, to re-searching the Web. Subjects appeared more willing to retain control over the number of re-search operations, but were willing to experiment with reordering. This suggests an association between the severity of the decision and subject's willingness to retain control over them. That is, for less severe strategies subjects were more willing to delegate responsibility to the system.

The implicit feedback frameworks evaluated in this thesis are dependent on how results are presented and how searchers interact with them. In the next section I discuss the presentation of information at the results interface and aspects of subject interaction.

12.3.4 Presentation and Interaction

In all experimental systems subjects suggested that they tried to look at information related to the search task. This was an important aspect of the experimental systems that used implicit feedback since they relied on subjects using the interface components as feedback on what information is relevant. It has been well documented that searchers will demonstrate a variety of information seeking behaviours during the course of a search (Ellis, 1989; Hancock-Beaulieu, 1990; Kuhlthau, 1991), and indeed will exhibit different kinds of interaction with different texts according to different goals, knowledge and intentions. However, searcher interaction is generally driven by a desire to maximise the amount of relevant information they view (maximise recall), whilst also minimising redundancy (maximise precision). Through monitoring the information they interact with I have shown that search systems can approximate subject's information needs.

The direct involvement of the searcher in the information seeking process results in a dialogue between them and the IR system that is potentially muddled and misdirected (Ingwersen, 1992). The systems described in the later parts of this thesis implement aspects of the principle of *polyrepresentation* (Ingwersen, 1994) that suggests one should provide and use different cognitive structures during acts of communication to reduce the uncertainty associated with interactive IR. The cognitive structures around which polyrepresentation is based are manifestations of human cognition, reflection or ideas. In IR the author's text, including titles and the full-text are representations of cognitive structures intended to be communicated. However, these portions of text demonstrate different functional origins. That is, they have the same cognitive origin but were created in a different way or for a different purpose. Subjects generally responded well to the content-rich interfaces and suggested that the multiple document representations allowed them to focus on the most relevant parts of the documents. Some subjects remarked that they would like to be able to jump between steps in a relevance path. For example, in the search interfaces presented in Chapter Ten a searcher cannot move straight from a top-ranking sentence to that sentence in its source document context. This rigidity of the relevance path structure is a necessity of the implicit feedback model deployed (which is path based). The Binary Voting Model, described in Chapter Six, does not place such constraints on path traversal and would perhaps be more suited for search interfaces that wish to implement a less rigid term weighting methodology.

Overall, the findings suggest that subjects want to retain control over the strategic aspects of their interaction. That is, over the aspects that will directly influence the quality of the results offered or future directions of their search. They view the provision of relevance indications only as an operational activity required to receive assistance. There is a disparity between how important subjects regard the communication of relevance information and its importance to the search system. Although relevance feedback can be useful tool to improve search effectiveness, it is under utilised because of the interface techniques it uses to gather relevance information. To cater for this, search systems must incorporate new techniques for gathering relevance information. Implicit relevance feedback methods such as those described in this thesis may be useful to address this problem. Further research is required in the development of search tools that incorporate implicit feedback techniques for gathering relevance information.

In the next section results relating to the next research question – the effectiveness of the information need detection component – are discussed.

12.4 Information Need Detection

Searchers may have problems choosing terms to adequately represent their information needs (Taylor, 1968). In this thesis approaches for choosing terms to create new, improved queries are presented and evaluated with human subjects and a novel simulation-based evaluation methodology. In this section I discuss experimental findings on the information need detection part of the implicit feedback framework. This experiment tested the term selection component of the framework from the subjects' perspective in a series of information seeking scenarios on different experimental systems. The simulation-based study in Chapter Eight allowed me to benchmark the performance of the term selection models with simulated searchers. The success of the Jeffrey's Conditioning Model meant it was selected to choose terms for query modification in this experiment.

The same model was used in three interfaces and differences in subjects' perceptions of the relevance and usefulness of the terms were noticed between systems. This suggests that the way the terms are presented plays an important part in how the terms are perceived, independent of their value. Subjects were asked to assess the 'relevance' and 'usefulness' of the terms suggested by the framework. In task-oriented evaluations one would expect relevance to be synonymous with 'utility' (Cooper, 1973) or 'pertinence' (Saracevic, 1996), resulting in a strong correlation between relevance and usefulness. However in the evaluation there were statistical differences between the relevance and usefulness scores for five of the six system-group comparisons and overall among all subjects and all systems; subjects generally regarded terms as being more relevant than useful. This could be because subjects did not know what relevance was or they did not associate it with usefulness. Five of the 48 subjects commented on the difference between relevance and usefulness; they could recognise which terms are related to the search (topically relevant) but not which were useful in pushing the search forward in terms of changing search focus or retrieving more relevant documents (useful). So although they can recognise easily that terms are on topic they may have trouble saying which were useful. This example demonstrates the importance of asking the right questions in user experiments such as this. There is a danger that experimenters would typically ask whether the terms selected by the system are 'relevant' or 'useful', but not both. In doing so they would miss the distinction one can make between the two attributes.

Subjects assessed the usefulness of terms on a five point semantic differential, between 1 and 5 (inclusive). The lower the score assigned the more useful the terms. Overall, across all systems and subjects, the terms chosen by the system were assigned an average score of 2.18. This score was worse than one, the lowest (best) possible value. In Pilot Test 1 subjects did not rate their own search terms as *always* useful, they acknowledge that they are not able to adequately conceptualise their information need, even when given the chance to refine the terms used to express it. However, as they view and process information, and their state of knowledge changes, they become more able to express these needs. The term selection model learns in a similar way, training itself with searcher interaction to better define what is relevant. It is difficult for any feedback model to choose useful terms, especially if subjects cannot even regard the terms they choose as useful. Unlike the discussion of interface support mechanisms in the previous section there were no differences in the usefulness of terms selected by the model for different types of search tasks.

Search systems that use implicit feedback techniques typically make decisions on behalf of searchers to assist them in their search. To operate effectively, such systems need to gain the trust of those that use them. In this experiment subjects were asked to indicate the extent to which they would trust the three experimental systems to choose terms on their behalf. The results again indicated a preference for the Recommendation system even though the same term selection model was used in all systems; both groups of subjects associated a higher level of trust with the Recommendation system. The true level of trust in the information need detection component is best measured independent of subject groups and independent of experimental systems. The average differential was 2.12, suggesting that subjects trusted the term selection component. The finding suggests that how the system communicates its decisions impacts on the level of trust subjects have in it.

During their searches subjects added new terms to their queries. These terms originated in ideas from a number of sources: (i) the terms recommended by the system, (ii) the retrieved documents and extracted information, (iii) a combination of these first two, (iv) the task being attempted, and (v) the subjects' tacit knowledge. The ideas derived from their search can result in a change in the direction of the search or the refinement of the current query statement with terms that better express information needs or better fit with the vocabulary of the collection. The terms suggested by all experimental systems appeared useful to initiate new ideas with around 20% of all new terms coming from ideas given by terms selected by the system. Ideas for terms also came from other sources, such as the task description, although it is conceivable that subjects will not always have search description as carefully constructed as a simulated work task situation.

The findings show that in systems that removed searcher control (i.e., the Automatic and Recommendation systems), subjects were more likely to use the terms proposed to initiate new ideas and search directions. The Checkbox system was dependent on subjects marking results as relevant, and as a consequence, the terms suggested were from items the subjects already knew were relevant. In situations where searchers may benefit from a change in search direction it may be better to gather feedback implicitly as this can provide insight into their general, rather than exact, interests. Systems that remove searcher control over creating queries may be most appropriate for encouraging new and potentially useful search directions. This can be helpful if the searcher is struggling with their search. Although the findings discussed in the previous section suggest that searchers want to retain control over the additional terms used, it may not be in their interests to do so, especially if they lack the experience to devise well-formed queries.

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

In the Recommendation system subjects were given a longer list of terms so they could be more selective about what terms were added. Subjects confirmed that the difference in the results was not related to the larger number of terms shown by the Recommendation system, but to the nature of the interface. Subjects were asked a simple 'yes'/'no' question as part of the informal discussion that followed the task on the Recommendation system. They were asked whether the larger number of terms in this system had an effect on their perceptions of the terms suggested; 42 of the 48 subjects responded 'no'; those that responded 'yes' found terms at a low-ranked position in the recommended list useful in their search. Subjects associated their preference for the Recommendation system with their perceptions of the query terms, showing that presentation factors can affect subject perceptions of such terms. In this experiment, the longer lists of suggested terms in the Recommendation system had only a minimal effect. The query length was restricted to a maximum of ten terms and the average initial query length across all systems, subjects and tasks was 2.86 terms.

In each of the experimental systems subjects were shown the terms the system had selected for them. In the Recommendation and Checkbox systems they were given the option to edit their query (i.e., add or remove terms). The results showed that in both systems subjects typically accepted around 65% of the top six terms offered to them; demonstrating the effectiveness of the information need detection component. The Recommendation system showed 20 terms to the subject and allowed subject to move terms from anywhere in this list into the new query. In the analysis the list was divided into four quartiles, each containing 5 terms (i.e., the same number as in the Checkbox system). The scrollable window was sized so that the top six terms were shown at any time. The results show that more than three-quarters of terms (76.29%) came from the first 10 terms offered by the system; showing that the term weighting estimated which terms subjects were interested in. There were differences between subject groups in the rank position of terms chosen from the recommended term list. Experienced subjects were more likely to accept terms that appeared lower down the ranking (in the range 11-15). This may be because these subjects are interested in pushing the search forward through changing search focus or retrieving more relevant documents. Terms lower down the ranking may not be completely relevant and may foster the generation of new ideas.

In the studies described in Part II the experimental systems did not display the revised query, only the effect of the retrieval strategy that used the query (e.g., the reordered list of Top-Ranking Sentences). Subjects in those studies suggested that it would beneficial to see the terms used to allow them to make better decisions about the decisions made by the systems. In this experiment and in Pilot Test 1 subjects were shown the effect of the retrieval strategy chosen by the system and the revised query it created. That is, the query and its construction became a more prominent part of the search process.

In this section the results relating to the information need detection component of the system. The results showed that subjects found the terms selected by the framework relevant and useful in their search and that they would trust the framework to select terms for them. The terms chosen by the framework played a part in helping subjects create new query statements or make search decisions. In the next section findings related to the third research question, about the effectiveness of the information need tracking component, are discussed.

12.5 Information Need Tracking

The dynamic nature of information needs has been well documented (Bates, 1989; Harter, 1992; Bruce, 1994). As the need evolves, becoming more understood by the searcher, the searcher's actions and strategies may also evolve and a retrieval system should be able to adapt dynamically to this change. As well as refining query statements, the probabilistic framework also provides a mechanism through which it can support such evolving searches. The traditional view of information seeking assumes a searcher's need it static and represented by a single query submitted at the start of the search session. However, it may well be dynamic and could change to reflect the information viewed by the searcher. As they view this information their knowledge changes and so does their problematic situation.

In situations where a searcher's need is ill-defined and liable to change, Bates (1989) among others (Ellis, 1989; Kuhlthau, 1993b) has argued that it may be beneficial to first explore the information space in a multidimensional way, allowing searchers to understand their information need more clearly. The classic model of the IR involves the retrieval of documents in response to a query devised and submitted by the searcher. RF is an example of an iterative process to improve a search system's representation of a static information need. That is, the need after a number of iterations is assumed to be the same as at the beginning of the search; the aim of relevance feedback is not to provide information that enables a change in the direction of the search. In situations where the information need is vague or uncertain, information that searchers encounter is more likely to give them new ideas and consequently new directions to follow (Belkin *et al.*, 1993). At each stage searchers do not just modify the search terms used in order to get a better match for a single query, rather the information need (as well as the search terms used) is continually shifting, to various degrees.

Berrypicking (Bates, 1989) is a technique where the information required to satisfy a query is the culmination of the knowledge gleaned from documents examined during the search session (Belkin, 2000). The interface techniques used in this experiment (especially in the Checkbox system) encourage an information seeking strategy similar to berrypicking. Rather than viewing the full-text of documents and refining their own queries, searchers visit a variety of document representations and receive support in their query refinement from experimental systems. The search interface presents many representations of the same document, biased towards the initial search request. The Recommendation and Automatic systems observe the information seeking behaviour of the searcher and use the evidence it gathers to better define information needs and cater for changes in these needs. The presentation strategies are manifestations of the berrypicking metaphor. The Checkbox system allows fragments of information to be directly stored by the subject and used for query refinement. The Recommendation and Automatic systems make inferences about all the information viewed and selects retrieval strategies to suit the estimated degree of change.

Through monitoring the information stored or viewed by searchers, the framework generates revised query statements. It is the differences between the system's estimation of the information need as it generates these statements and its formulation near the beginning of the search that it uses to estimate the extent to which the need has changed. The framework chooses between three possible strategies aimed to support the user as they search; re-searching, reordering the document list, reordering the top-ranking sentence list and no-action at all. The strategies decrease in severity and reflect the estimated degree of change. Re-searching constructs a new information space and reordering restructures retrieved information depending on the level of change.

All subjects were instructed before the experiment that the different strategies provided varying degrees of interface support and had an increasingly dramatic effect on reshaping the information space. They were not told that the control related in any way to shifts, changes or developments in their information need as I felt this may bias their perceptions of the component. Searchers adapted well to the need tracking, and seemed comfortable with choosing between the different retrieval strategies.

The Recommendation and Automatic systems chose or recommended retrieval strategies. They were asked whether the retrieval strategy the system selected reflected any changes in their information need. There was a relationship between subject responses and the task categorisation used in this experiment. In the high complexity task there was scope for change whereas in the low complexity search task there was little.

The low complexity task was encouraged relevant or focused information seeking; the high complexity task encouraged explorative or browsing behaviour. Although the underlying topic is the same the additional restrictions placed on the low complexity search make the propensity to elicit changes in the information a subject is looking for also lower. [46] The findings of the experiment suggest that the information need tracking component was effective for high complexity tasks. The experimental systems selected more retrieval strategies for these types of task than for the low complexity tasks since in tasks of lower complexity subject's information needs remained more or less constant throughout their whole search. The more complex the search task, the more support subjects required in making decisions that had a strategic impact on their search. The information need tracking component appeared to not only track changes in the information needs, but the frequency of detected changes (and severity of chosen retrieval strategy) could be used to measure task complexity. For example, the selection of many retrieval strategies by the system may suggest that the search is variable and the search task is complex.

[46] The high complexity task is unclear about what information is being sought, how to obtain relevant information and how subjects will know when they have found relevant information. In contrast the low complexity task is generally clear about what information is required, how to find information and how to assess relevance.

Subjects were asked to rate how much they trusted the Recommendation and Automatic systems to select retrieval strategies for them. Although the subjects reacted positively to the retrieval strategy selected (i.e., the overall Likert scale response was significantly less than the middle value of the scale), they did not trust the information need tracking component as much as the information need detection component. This is perhaps because the potential implications of trusting the system to re-search information repositories or restructure the information displayed at the search interface are more severe than the selection of the some erroneous terms. Inexperienced subjects trusted the Automatic system less than the Recommendation system and since both used the same approach the difference may only be attributable to the presentation of the strategy. The Automatic system removed more control than the Recommendation system, selecting action and executing them without searcher consent. Inexperienced subjects commented that they did not feel in control of their search on the Automatic system. Experienced subjects felt similarly although some remarked that the removal of control was also a removal of burden and make the search simpler.

In a similar way to systems such as I^{3}R (Croft and Thompson, 1987) and FIRE (Brajnik *et al.*, 1996), the experimental systems created for the experiment in Part IV are always distinctly subordinate to the searcher. That is, the searcher always has the option to reverse system decisions. In Pilot Test 1 a search interface similar to that used in these experiments gave subjects the option to accept or reject search decisions after they occurred. In that experiment subjects commented that the communication of acceptance should be implicit as there was no need to tell the system they were happy with its decisions. In light of these comments a design decision was taken in the development of later experimental systems to only provide subjects with the option to 'undo'. Interaction logs were used to analyse the proportion of occasions that subjects reversed system decisions; around 70% of the search decisions made by the systems were accepted by subjects.

Some subjects commented that they would have liked to be shown a more comprehensive history of their search activity during their search including retrieval strategies chosen, queries submitted and all search results considered to be relevant. They also commented that they would like to be able to undo more than the previous action. In contrast, the experimental systems described in Part II did not provide any explicit notifications that search decisions had been made by the system or the option to reverse these decisions. In these systems a change in the rank order of the Top-Ranking Sentences was the first, and only, indication that the system had made a decision.

I measured the amount of overlap between the strategy chosen by the Checkbox and Recommendation systems and the strategies chosen by experimental subjects. A good information need tracking component should be able to predict searcher's decisions based on the variability of their search. This is a potentially difficult task and the reported success rate of 59.73% (almost 67% in the Recommendation system) appears reasonable. This was improved to 85.48% if I allowed some margin of error to include the search decision made and the next nearest decision (e.g., reorder Top-Ranking Sentences *and* reorder documents or reorder Top-Ranking Sentences *and* no action). The information need tracking component appears to make decisions that are appropriate for subjects' searches. In the next section I summarise the discussion presented in this chapter.

12.6 Chapter Summary

The results of the experiment show that it is possible to get searchers to interact with more than a few search results. The approach moves away from simply presenting titles to presenting alternative access methods for assessing and targeting potentially relevant information. From observations and informal post-search interviews across a series of related studies, subjects appeared to find the increased level of content shown at the results interface of value in their search. This is important, as the success of all experimental systems presented in this thesis – especially those that used implicit feedback techniques – is dependent on the use of these interface features.

The experiment tested different techniques for communicating relevance, creating queries and using these queries in different ways. Three experimental systems were developed that varied levels of control over each of these search aspects. These systems investigated which activities subjects wished to retain control over, and how much control they actually required. The results showed that searchers are happy to delegate full responsibility for indicating which search results are relevant, but only want to receive assistance in the formulation of query statements and selecting interactive search strategies. Subjects still wish to retain control over search activities they regard as important to the effectiveness of their search. Rather than trying to force searchers to provide feedback, implicit feedback techniques can remove the burden of indicating relevance, allowing subjects to focus on those activities they regard as important.

I found that the task categories used in the experiments were identifiable by subjects. That is, the variations in the task complexity were noticed by subjects even though they were not told that the complexity of the tasks differed. Subjects preferred the Recommendation system and found it better for more complex tasks where more control over the query terms was preferable. The Checkbox system was good for the low complexity tasks where the objective of the search was clear. The Automatic system was good for complex searches where the subject did not want to be actively engaged in the information seeking process or may lack insufficient knowledge about the retrieval environment to choose the good terms. In general, the systems communicated with searchers in a way that was helpful.

The terms selected for query modification were both useful and relevant. Subjects did not correlate relevance with usefulness suggesting that they interpreted them as being two different things in their search. The approach tracked potential changes or developments in the information need based on changes in the document representations viewed by the searcher. The system communicated its prediction of these changes through the search decisions it made on the subjects' behalf. The retrieval strategies chosen by the system were appropriate and liked by subjects.

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

In Part V I present the conclusions drawn from the research presented in this thesis and avenues for future work.

Part V

Conclusion

In Part IV I described a user experiment to test the performance of the information need detection and information need tracking components of the probabilistic framework. The results showed that the framework chose terms and strategies that were apt and liked by experimental subjects. The evaluation also compared experimental systems that varied the amount of control subjects had over conveying relevance information, creating new query statements and deciding how to use these new statements (i.e., how they interacted with the framework and it with them). The evaluation also showed that the subjects preferred implicit relevance indications to explicit and a system that made recommendations about additional terms and strategies over systems offering intrusive forms of support (where systems act directly) or passive forms of support (where systems await searcher action). In this part I conclude this thesis and present avenues for future work; both drawn from findings obtained in the user experiment and research described throughout this thesis.

Chapter 13

Conclusions

13.1 Introduction

In this thesis I have investigated the use implicit feedback techniques to help searchers use search systems more effectively. The components introduced help searchers create new queries and help them make new search decisions about how to use these queries to find new documents or reorganise information already retrieved. In Part II I described techniques to help searchers maximise the amount of useful information they can access during a search. In Part III, heuristic-based and probabilistic implicit feedback frameworks were introduced that use this interaction to revise queries and make search decisions. The term selection parts of these frameworks (and other baselines) were evaluated with a simulation-based evaluation methodology I devised to test how well each term selection model 'learned' relevance and improved search effectiveness The findings of Parts II and III motivate the development of the interfaces described in Part IV, where I present an investigation of how the implicit feedback framework should communicate with the searcher and vice versa. In this chapter I conclude and summarise the main findings and contributions of this thesis.

13.2 Content-Driven Information Seeking

In Part II I introduced new interface techniques to encourage searchers to search effectively by providing them with more information to make their decisions; I called this approach content-driven information seeking (CDIS). Unlike traditional result presentation techniques used by Web search engines such as Google, this approach shifts the focus of interaction at the results interface from documents to the information resident inside documents. To do this it uses query-relevant *Top-Ranking Sentences* extracted from top documents as an interface component to facilitate effective information access. Top-Ranking Sentences are a precision-oriented approach I devised to maximise the amount of useful information a searcher can access. I conducted three related user studies to test the effectiveness of these sentences with real searchers in different search scenarios. In the first study, I used the ranked sentences as an alternative to document lists, shifting searcher attention from the document surrogates (i.e., titles, sentence fragments and URLs) to document content. The second used the sentences to reflect the use of two contrasting relevance feedback techniques. The third used the sentences to encourage interaction with the retrieved set, to reflect the dynamic nature of the information need and to complement, rather than replace, document lists. Each study involved human subjects and different types of information seeking scenario based around simulated work task situations. I showed that the CDIS approach, whether or not supported by additional implicit feedback techniques that reorder the sentences, can lead to effective and efficient searching.

As part of the exposition of CDIS, I also introduced the notion of 'push' and 'pull' information seeking and explained that these approaches differ in how information is presented to the searcher. Motivated by the success of the techniques in the studies described in Chapter Four, I extended the CDIS approaches in Chapter Five with the inclusion of more document representations and relevance paths that join them. Content-rich search interfaces were developed using these additional representations to encourage interaction and create more evidence for the implicit feedback frameworks introduced in Part III. In user studies of interfaces that used these additional components (Pilot Test 1 and presented in Part IV) I showed that searchers found them helpful, that they encouraged more interaction with search results and that they felt the additional interaction was beneficial to them; this benefit was more apparent when information needs were vague or the search tasks complex.

Ranking documents is a cumbersome means of result presentation. Documents may not be entirely relevant and document titles, sentence fragments and URLs may not be strictly indicative; it is the information inside documents that searchers generally seek. The CDIS approach I introduced extracts, ranks and presents potentially relevant content from the returned set, blurring inter-document boundaries and encouraging information seeking based on the pertinent document content. In the next section I describe implicit feedback frameworks that use interaction with content-rich search interfaces as evidence to help them make search decisions.

13.3 Implicit Feedback Frameworks

In Part III two novel implicit feedback frameworks were introduced: one heuristic-based and one probabilistic. The frameworks estimate current information needs and estimate changes in those needs as a searcher interacts with the results of their retrieval. The frameworks presented support searchers by passively observing their search behaviour and choosing new query terms and retrieval strategies to help them locate relevant information. They aim to help those who are unable or unwilling to communicate relevance information directly or simply may be struggling to find what they want.

Motivated by the success of interface components and the implicit feedback techniques described in Part II the implicit feedback frameworks I created approximate searcher interests through interaction with representations of top-ranked documents and interactive paths that join them. This differs from traditional potentially unreliable sources of implicit feedback such as document reading time, scrolling and other such measurable search behaviours within the full-text of potentially relevant documents. In rich information seeking environments like those created by CDIS techniques, searchers can view information to a fine level of detail and the information they view can be used to approximate their interests. The frameworks performed this function well and provided a means through which searcher intentions could be inferred implicitly, without the need for direct searcher involvement in providing relevance information.

As I established in Part I, information needs are not static and can change during a search on exposure to new information. The implicit feedback frameworks contain components that allow them to predict when, and by how much, the topic of a search has changed based on short-term, within search session, interaction histories. Depending on the degree of the change the frameworks can pick retrieval strategies that will be useful to searchers. That is, the level of interface support offered by systems that implement these frameworks depends on the extent to which information needs are estimated to change. There are four possible strategies the framework can follow: no action (for small changes), reorder top-ranking sentence list (for small-moderate changes), reorder document list (for moderate-large changes) and re-search (for large changes). I conducted a study of topic similarity measures that demonstrated the effectiveness of correlation coefficients for predicting the extent of the difference between search topics. The results show that measures based on the level of correlation between topics concords highly with general subject perceptions of search topic similarity and that these coefficients may be useful to predict search topic change. As a result, the Spearman and Pearson correlation coefficients were used as tools to estimate changes in searcher interests and select appropriate retrieval strategies.

Two user experiments involving a total of 72 different subjects (i.e., Pilot Test 1 and the experiment in Part IV) have shown that the heuristic-based and probabilistic implicit feedback frameworks choose new query terms and make decisions about query use that are appropriate and liked by experimental subjects. The techniques discussed in this thesis have the potential to alleviate some of the problems inherent in traditional RF – where searchers are directly involved in the provision of relevance information – whilst preserving the benefits that underlie the approach. The initial query is still modified to become attuned to a searcher's need based on an iterative process of feedback. However, searchers do not have to explicitly assess and mark documents as relevant and the way the new query is used depends on the extent to which the information need is estimated to have changed (i.e., the search systems do not only re-search the document collection). In the next section I describe the simulation-based evaluation methodology I developed to test the term selection models that in the implicit feedback frameworks. This methodology is used as a formative evaluation technique to select the best-performing model for implementation in the search interfaces described in Part IV.

13.4 Simulation-Based Evaluation Methodology

A novel simulation-based evaluation methodology was used to test the performance of the term selection components of implicit feedback frameworks (called implicit feedback models) in different simulated contexts. This methodology is less time consuming and costly than experimentation with human subjects, allows environmental and situational variables to be more strictly controlled and complex searcher interactions to be modelled. It allowed me to compare and fine-tune a number of potential implicit feedback models before the best performing model was deployed in an interactive search system. Simulations of this nature could be a powerful formative evaluation tool for the designers of search interfaces, especially those that do not conform to traditional forms of search result presentation (i.e., ranked lists of documents). Designers can test a prototype interface with one implicit feedback model to remove potentially problematic interactions or, as I have described in this thesis, test many models for a given search interface to choose the most effective model.

The implicit feedback models tested were ostensive in nature and use the exploration of the retrieved information and the viewing of document representations as an indication of relevance. Six implicit feedback models were tested in total, all using an ostensive paradigm but each employing a different term selection stratagem. The methodology tested those models in different search situations.

I introduced implicit feedback models based on Jeffrey's rule of conditioning, Binary Voting, three that use the popular *wpq* query expansion approach and a baseline that selected terms randomly. The simulated approach used to test the models assumes the role of a searcher 'viewing' relevant documents and relevance paths between different representations of documents. The simulation passes the information it viewed to the implicit feedback models, which use this as evidence of relevance to select terms to best describe this information. In the evaluation I investigated the degree to which each of the models improved search effectiveness and 'learned' what information was relevant. From the six implicit feedback models tested, the Jeffrey's Conditioning Model was most effective. As demonstrated in Chapter Eight, this model outperformed the others in a variety of different simulated search scenarios with different proportions of relevant and non-relevant information and other interaction constraints. This model was subsequently chosen as the term selection component of the implicit feedback framework tested in the experiment described in Part IV. During this experiment the ability of the model to identify information needs was re-tested with human subjects. The results of that experiment showed that the model chose terms that were relevant and useful.

The simulation-based evaluation methodology I propose is an effective way of testing the worth of implicit feedback models such as those presented in this thesis. Experimentation with human subjects can be costly and these tests can ensure that only the best models are chosen to be tested with real searchers in interactive information seeking environments. The next section discusses issues in the interface support offered to searchers.

13.5 Interface Support

The results of all user experiments described in Parts II and IV show that it is possible to get searchers to interact with more than a few search results. The approaches introduced move away from simply presenting titles to presenting alternative access methods for assessing and targeting potentially relevant information. From observations and informal post-search interviews across a series of related studies, subjects appeared to find the increased level of content shown at the results interfaces of value in their search. This is important, as the success of all experimental systems I present – especially those based on implicit feedback techniques – is dependent on the use of these interface features.

In Part IV of this thesis I investigated how implicit feedback frameworks can best communicate with searchers (and vice versa) and evaluated the implicit feedback framework chosen from the findings of the simulations in Part III. The experiment used three different types of RF interface that varied how searchers provided relevance information, how they created new queries and how they made new search decisions. Three systems were created: a Checkbox system that relied on explicit relevance assessments, provided support in creating queries and relied on searcher to select retrieval strategies; a Recommendation system, that gathered implicit relevance assessments and *recommended* query terms and strategies, and an Automatic system that gathered implicit relevance assessments and *selected* terms and strategies.

In this experiment, subjects preferred the Recommendation system and found it useful for more complex tasks where more control over the query terms was preferable. Subjects found the Checkbox system useful for low complexity tasks where the objective of the search was clear. Subjects found the Automatic system useful for complex searches where the subject did not want to be actively engaged in the information seeking process or lacked sufficient knowledge about the retrieval environment to make good decisions. The different systems were therefore useful for different types of search, although the Recommendation system (originally devised based on the feedback of subjects in Pilot Test 1) was generally most popular as it gave searchers control over query term selection and use.

The terms selected for query modification by the probabilistic framework were both useful and relevant and were accepted by searchers on different systems since they were valuable for their search tasks. The approach tracked potential changes or developments in the information need based on changes in the document representations viewed by the subject. The system communicated its estimation of these developments through the decisions it made on subjects' behalf; subjects generally felt these strategies were useful and appropriate.

Implicit relevance information is inherently uncertain. The Recommendation system worked in tandem with the searcher, making suggestions on what terms they could add or what strategies they could select. The uncertainty surrounding how implicit evidence is gathered means that it is desirable to give searchers final control over systems that use it. In the experiment in Part IV (as in Pilot Test 1) subjects wished to retain control over activities they perceived as being important for the success of their search. That is, subjects were willing to delegate control over the provision of relevance information (i.e., the inputs) as long as they could control how this information was used in constructing new queries or making new search decisions (i.e., the outputs).

13.6 Chapter Summary

In this thesis I have presented and evaluated a set of techniques to support searchers engaged in interactive information retrieval. I have developed novel search interfaces that 'push' potentially relevant information toward the searcher, helping them proactively as they search. I have developed content-rich search interfaces that extend this approach to involve a greater variety of document representations and interactive relevance paths that join these representations. These interface techniques have been shown to help searchers, especially for complex search tasks. I have developed and tested heuristic-based and probabilistic implicit feedback frameworks that use interaction with these content-rich interfaces to estimate and track information needs. User experiments have shown that the frameworks select terms and retrieval strategies that subjects found appropriate and helpful.

I developed a simulation-based evaluation methodology for testing implicit feedback models with simulated searchers and benchmarked the performance of six different models in a variety of retrieval scenarios. The methodology allows complex interaction to be modelled and experimental variables to be closely controlled whilst giving system designers a formative evaluation tool to assist in the selection of RF algorithms or design of search interfaces. The best model was chosen to be part of an experiment to further test it with human subjects and with different types of interface support in feedback systems. I developed three RF search interfaces, whose design was motivated by findings in my earlier studies, each using the best performing model and each with different interface options that afforded different amounts of searcher control. The results showed that searchers are happy to delegate responsibility for relevance assessment to RF systems (through implicit feedback), but not more severe decisions such as formulating queries or selecting retrieval strategies; for such decisions searchers wanted support from the system, but ultimately control over its actions.

This research has investigated innovative techniques for interface design, implicit feedback and evaluation for interactive IR. The ramifications of this work are notable and warrant further investigation. The final chapter will outline potential avenues for such investigation in future work.

Chapter 14

Future Work

14.1 Introduction

This thesis has explored many issues in the areas of implicit feedback and interactive information retrieval. Many avenues have emerged for the research described to be taken further and in this chapter I describe some of the main opportunities and challenges that this work provides. In the same way as the previous chapter I discuss future work in a number of sections, based on the contributions made by this thesis.

14.2 Content-Driven Information Seeking

The presentation of multiple representations of search results at the results interface was promising and liked by searchers in all user studies conducted. There were minor issues with the presentation of this content, such as the occlusion of other information when viewing document summaries or sentences in context, although these could be resolved with slight modifications to interface design. The results of the experiments in Chapter Four suggested that the content-driven approaches were of most use for search tasks where a lot of information is preferred to improve topic familiarity or awareness (*background* search) or improve decision making abilities (*decision* search). However, the approach was not of as much use for 'fact' searches where the information need was exact. Since the content-driven approaches are not as effective in all information seeking contexts it is important to identify when the approach should be used and when it should not. That is, when should a searcher be presented with a list of Top-Ranking Sentences and other interface components, and when should they be faced with a ranked list of document surrogates. The decision of when to use content-driven information seeking techniques should ideally be taken by the search system, since searchers may not realise the potential benefits of the approach. There is future work in developing mechanisms to make these decisions.

As suggested in Chapter Five, the interfaces described in this thesis implement aspects of a polyrepresentative approach i.e., presenting multiple document representations to reduce uncertainty in implicit feedback. In future work I will analyse the interactive experiments conducted with these interfaces in Pilot Test 1 and in the experiment presented in Part IV from the perspective of polyrepresentation and use the system logs of mouse movements and clicks to allow me to better understand and interpret system usage.

14.3 Implicit Feedback Frameworks

The implicit feedback frameworks chose additional search terms that were relevant and useful for searchers. However, a larger scale empirical evaluation to improve the indicativity heuristics used in the Binary Voting Model may improve the effectiveness of the heuristic-based framework. The performance of both frameworks could also be improved if searchers could indicate what information is definitely not relevant. During the experiments some subjects suggested that they wanted control over what information the search system disregarded and excluded from the search. The issues about whether searchers are actually able to exercise the control to provide negative RF effectively has already been raised by Belkin *et al.* (1998). Potential avenues for future work could be on the development of hybrid positive/negative implicit/explicit RF systems that gather positive assessments unobtrusively and negative assessments directly from the searcher. In this thesis searcher actions such as regressing back along a relevance path, or reversing a search decision made by the search system were ignored by the frameworks and only positive assessments were considered. Further work is needed in using these indications of dissatisfaction to infer what information searchers do not want.

The results of Pilot Test 1 and the experiment presented in Part IV suggests that the weakest part of the implicit feedback frameworks is the component to estimate changes in information needs. Developing sound techniques to track changing needs can be difficult as searchers may be unaware that any change has occurred and needs may change in different ways to different degrees. However, in future work there is a need to address the shortfall between searcher expectations of the component and its actual performance.

The frameworks currently only track information needs during a single search session. An avenue for future work would be test the effect of incorporating searchers' long term interests. These interests could be used to develop a potentially more robust formulation of the information need from which query terms could be chosen. The frameworks I have introduced use interaction with IR system result interfaces as implicit feedback. A deeper understanding of how searchers interact with the full-text of relevant documents is needed before traditional implicit feedback metrics (e.g., viewing time or scrolling) could be used to complement these frameworks I present here.

14.4 Simulation-Based Evaluation Methodology

The simulation-based evaluation methodology allowed feedback models to be compared in an experimental setting without human subjects. Simulations of this nature can be used either after a prototype interface was built (as was the case in this thesis), or before the interface is built, to test its performance with every possible set of potential searcher interactions prior to development. Testing of this nature can assist system designers in identifying the strengths and weaknesses of the interface with a particular implicit feedback model (allowing them to eliminate interactions that could cause problems) or the strengths and weaknesses of many implicit feedback models for a given search interface (allowing them to choose a model that suits their needs). More work is necessary in developing a framework to allow simulations of this nature of be developed in a robust, generic and extensible way.

To more closely emulate the search behaviour of humans the simulations need to make decisions that resemble those that human subjects may make. In future work I will address this issue and try to make the methodology more 'intelligent' to allow better [47] decisions on what information to interact with and develop a suite of simulated searchers, each with their own stereotypical search behaviour. To test a model or system interface with, for example, experienced searchers, it should be possible to select a group of simulations with the appropriate characteristics, and plug them into the methodology. There is much scope for future work in developing effective searcher simulations to model different scenarios, searcher and searching style.

Searcher simulations could also be used to mimic changes in the topic of the search and monitor how well the relevance feedback techniques adapt to this change and how well components to track changes in information needs detect these changes. Changes in the search topic are potentially difficult to estimate as information need change or development is perhaps more difficult to monitor than the information need itself, which can be approximated at the relevant document level or through decent query terms. More work is necessary in simulating different rates of change and different search strategies and tactics used during this change.

[47] To simulate novice searchers or those engaged in complex search tasks the methodology may also need to make bad decisions.

14.5 Interface Support

In Part IV I tested three experimental interfaces that gave searchers different levels of control and responsibility over aspects of the search. The interfaces were generally liked by searchers, and while they were happy to delegate responsibility for gathering relevance information to the search system they wished to retain control of query creation and retrieval strategy selection. This suggested that subjects did not trust the implicit feedback framework sufficiently to give it complete control over all search decisions. More work is therefore necessary to engender trust in system decisions by improving the effectiveness of the framework and how the decisions are communicated at the search interface. The use of explanations, such as that proposed by Ruthven (2002), may help searchers understand why certain terms were chosen and search decisions made. Further work is necessary to investigate task and situational differences in searcher control and responsibility.

Further work is also necessary on the association between the support offered by the experimental system and the complexity of the search task. The testing of interfaces in laboratory settings may not reveal problems encountered in operational settings, where IR systems are typically used. In future work a longitudinal evaluation of the systems in an operational environment is essential to test the worth of the interface approaches proposed.

14.6 Chapter Summary

This chapter has detailed opportunities to further the research presented in this thesis. The techniques proposed have fundamental implications for the design of interactive information retrieval systems and their evaluation. This work has shown that searchers respond well to the content-driven information seeking approaches and the implicit feedback frameworks that use them. The simulation-based evaluation techniques I propose provide a means through which interfaces and their underlying mechanisms can be assessed. It is vital therefore that more work is undertaken to further this imaginative research, and test these concepts in operational environments and longitudinal user experiments.

References

Aalbersberg, I. J. (1992). Incremental relevance feedback. *Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 11-22).

Allan, J. (1995). Relevance feedback with too much data. *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 337-343).

Allan, J. (1996). Incremental relevance feedback for information filtering. *Proceedings of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 270-278).

Allen, R. B. (1990). User models: Theory, method and practice. *International Journal of Man-Machine Studies*, 32, 511-543.

Amitay, E. and Paris, C. (2000). Automatically summarising web sites: Is there a way around it?. *Proceedings of the 9th International Conference on Information and Knowledge Management*. (pp. 173-179).

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

Anick, P. and Tipirneni, S. (1999). The paraphrase search assistant: Terminological feedback for iterative information seeking. *Proceedings of the 22nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 153-159).

Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T. (1995). WebWatcher: A learning apprentice for the world wide web. *Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*. (pp. 6-12).

Azman, A. and Ounis, I. (2004). Discovery of aggregate usage profiles based on clustering information needs. *Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 470-471).

Balabanovic, M. and Shoham, Y. (1995). Learning information retrieval agents: Experiments with automated web browsing. *Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*. (pp. 13-18).

Ball, G. and Breese, J. (1999). Modelling the emotional state of computer users. *Proceedings of the Workshop on Personality and Emotion in User Modelling*.

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

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

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

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370-418.

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

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

Beaulieu, M., Robertson, S. E. and Rasmussen, E. (1996). Evaluating interactive systems in TREC. *Journal of the American Society for Information Science*, 47 (1), 85-94.

Belkin, N. J. (1980). Anomalous state of knowledge for information retrieval. *Canadian Journal of Information Science*, 5, 133-143.

Belkin, N. J. (1984). Cognitive models and information transfer. *Social Science Information Studies*, 4, 111-129.

Belkin, N. J. (2000). Helping people find what they don't know. *Communications of the ACM*, 43 (8), 59-61.

Belkin, N. J., Brooks, H. M. and Daniels, P. J. (1987). Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studies*, 27, 127-144.

Belkin, N. J., Cool, C., Head, J., Jeng, J., Kelly, D., Lin, S., et al. (2000). Relevance feedback versus local context analysis as term-suggestion devices: Rutgers' TREC-8 interactive track experience. *Proceedings of the Eighth Text Retrieval Conference*. (pp. 565-574).

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

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

Belkin, N. J., Cool, C. and Koenemann, J. (1996a). On the potential utility of negative relevance feedback for interactive information retrieval. *Proceedings of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 341).

Belkin, N. J., Cool, C., Koenemann, J., Bor Ng, K. and Park, S. Y. (1996b). Using relevance feedback and ranking in interactive searching. *Proceedings of the Fourth Text Retrieval Conference*. (pp. 181-210).

Belkin, N. J., Cool, C., Stein, A. and Theil, U. (1993). Cases, scripts and information seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 29 (3), 325-344.

Belkin, N. J. and Croft, W. B. (1992). Information filtering and retrieval: Two sides of the same coin? *Communications of the ACM*, 35 (12), 29-38.

Belkin, N. J., Oddy, R. N. and Brooks, H. M. (1982). ASK for information retrieval: Part I - background and theory. *Journal of Documentation*, 38 (2), 61-71.

Belkin, N. J., Perez Carballo, J., Cool, C., Lin, S., Park, S. Y., Rieh, S. Y., et al. (1998). Rutgers TREC-6 Interactive Track Experience. *Proceedings of the 6th Text Retrieval Conference*, 597-610.

Belkin, N. J. and Vickery, A. (1985). *Interaction in information system: A review of research from document retrieval to knowledge-based system*, 188-198. London: The British Library.

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

Berger, A. L. and Mittal, V. O. (2000). OCELOT: A system for summarizing web pages. *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 144-151).

Billsus, D. and Pazzani, M. J. (1999). A personal newsagent that talks, learns and explains. *Proceedings of the 3rd International Conference on Autonomous Agents*. (pp. 268-275).

Borlund, P. (2000a). *Evaluation of interactive information retrieval systems*. Unpublished doctoral dissertation, Åbo Akademi University,

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

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

Borlund, P. and Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53 (3), 225-250.

Borlund, P. and Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: performance indicators for interactive IR. *Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 324-331).

Brajnik, G., Guida, G. and Tasso, C. (1987). User modeling in intelligent information retrieval. *Information Processing and Management*, 23 (4), 305-320.

Brajnik, G., Mizzaro, S. and Tasso, C. (1996). Evaluating user interfaces to information retrieval systems: A case study of user support. *Proceedings of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 128-136).

Brandow, R., Mitze, K. and Rau, L. F. (1995). Automatic condensation of electronic publications be sentence selection. *Information Processing and Management*, 31 (5), 675-685.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36 (2), 3-10.

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

Bruza, P., McArthur, R. and Dennis, S. (2000). Interactive internet search: Keyword, directory and query reformulation mechanisms compared. *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 280-287).

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

Budzik, J. and Hammond, K. J. (2000). User interactions with everyday applications as context for just-in-time information access. *Proceedings of the Annual Conference on Intelligent User Interfaces*. (pp. 44-51).

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

Byström, K. and Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, 31 (2), 191-213.

Callan, J. P. (1994). Passage-level evidence in document retrieval. *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 302-310).

Campbell, D. (1988). Task complexity: a review and analysis. *Academy of Management Review*, 13, 40-52.

Campbell, I. (1999). Interactive evaluation of the ostensive model, using a new test collection of images with multiple relevance assessments. *Journal of Information Retrieval*, 2 (1), 89-114.

Campbell, I. (2000). *The ostensive model of developing information needs*. Unpublished doctoral dissertation, University of Glasgow, Glasgow.

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

Chalmers, M. and Chitson, P. (1992). Bead: Explorations in information visualisation. *Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 330-337).

Chalmers, M., Rodden, K. and Brodbeck, D. (1998). The order of things: Activity-centred information access. *Computer Networks and ISDN Systems*, 30 (1-7), 359-367.

Chen, H. and Dumais, S. T. (2000). Bringing order to the web: Automatically categorizing search results. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. (pp. 145-152).

Choo, C. W., Detlor, B. and D., T. (1999). Information seeking on the web: An integrated model of browsing and searching. *FirstMonday*, 5 (2), 3-.

Claypool, M., Le, P., Waseda, M. and Brown, D. (2001). Implicit interest indicators. *Proceedings of the 6th International Conference on Intelligent User Interfaces*. (pp. 33-40).

Cleverdon, C. W. (1960) *Aslib Cranfield research project: Report on the first stage of an investigation into the comparative efficiency of indexing systems*. Cranfield: The College of Aeronautics.

Cooper, M. D. and Chen, H.-M. (2001). Predicting the relevance of a library catalog search. *Journal of the American Society for Information Science*, 52 (10), 813-827.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7 (1), 19-37.

Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness: Part I. *Journal of the American Society for Information Science*, 24, 87-100.

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). I3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, 38 (6), 389-404.

Crouch, C. J., Crouch, D. B., Chen, Q. and Holtz, S. J. (2002). Improving the retrieval effectiveness of very short queries. *Information Processing and Management*, 38 (1), 1-36.

Cutting, D., Karger, D., Pedersen, J. and Tukey, J. W. (1992). Scatter/Gather: A cluster-based approach to browsing large document collections. *Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 318-329).

Dennis, S., McArthur, R. and Bruza, P. (1998). Searching the world wide web made easy? The cognitive load imposed by query refinement mechanisms. *Proceedings of the Third Australian Document Computing Symposium*. (pp. 65-71).

Driori, O. (2003). How to display search results in a digital libraries-user study. *Proceedings of the 3rd Workshop in New Developments in Digital Libraries*. (pp. 13-28).

Dumais, S. T., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R. and Robbins, D. C. (2003). Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. *Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 72-79).

Dumais, S. T., Cutrell, E. and Chen, H. (2001). Optimizing search by showing results in context. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. (pp. 277-284).

Dziadosz, S. and Chandrasekar, R. (2002). Do Thumbnail Previews Help Users Make Better Relevance Decisions about Web Search Results? *Proceedings of 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 355-356.).

Earl, L. L. (1970). Experiments in automatic extracting and indexing. *Information Storage and Retrieval*, 6 (6), 313-334.

Edmundson, H. P. (1964). Problems in automatic abstracting. *Communications of the ACM*, 7 (4), 259-285.

Edmundson, H. P. (1969). New methods in automatic abstracting. *Journal of the ACM*, 16 (2), 264-285.

Efthimiadis, E. N. (1993). A user-centred evaluation of ranking algorithms for interactive query expansion. *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 146-159).

Eisenberg, M. and Barry, C. L. (1988). Order effects: a study of the possible influence of presentation order on user judgements of document relevance. *Journal of the American Society for Information Science*, 39 (5), 293-300.

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation*, 45, 171-212.

Fendlay, J. M., Walker, R. and Kengridge, R. W. (Eds.). (1995). *Eye movement research: Mechanisms, processes and applications*. New York: Elsevier Science Publishing.

Fischer, G. (2000). User Modeling in Human-Computer Interaction. *User Modeling and User-Adapted Interaction*, 11 (1-2), 65-86.

Florance, V. and Marchionini, G. (1995). Information processing in the context of medical care. *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 158-163).

Ford, N. (1980). Relating information needs to learner characteristics in higher education. *Journal of Documentation*, 36, 165-191.

Fowkes, H. and Beaulieu, M. (2000). Interactive searching behaviour: Okapi experiment for TREC-8. *Proceedings of the 22nd BCS-IRSG European Colloquium on IR Research*.

Furnas, G. W. (1997). Effective view navigation. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computer Systems*. (pp. 367-374).

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

Furnas, J. (2002). On recommending. *Journal of the American Society of Information Science and Technology*, 53 (9), 747-763.

Gauch, S. (1992). Evaluation of an expert system for searching in full text. *Proceedings of the 13th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 255-257).

Gemmel, J., Bell, G., Lueder, R. and Gelernter, D. (2002). MyLifeBits: Fulfilling the MEMEX vision. *Proceedings of ACM Multimedia*. (pp. 235-238).

Goldstein, J., Mittal, V. O., Carbonell, J. and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*.

Golovchinsky, G. (1997). What the query told the link: The integration of hypertext and information retrieval. *Proceedings of the 8th ACM Conference on Hypertext*. (pp. 67-74).

Gong, Y. and Lui, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. *Proceedings of 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 19-25).

Hagerty, K. (1967). *Abstracts as a basis for relevance judgment*. Unpublished doctoral dissertation, University of Chicago, Chicago.

Hamming, R. W. (1950). Error-detecting and error-correcting codes. *Bell Systems Technical Journal*, 29, 147-160.

Hancock-Beaulieu, M. (1990). Evaluating the impact of an online library catalogue in subject searching behaviour at the catalogue and at the shelves. *Journal of Documentation*, 46, 318-338.

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

Harman, D. (1986). An experimental study of the factors important in document ranking. *Proceedings of the 9th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 186-193).

Harman, D. (1988). Towards interactive query expansion. *Proceedings of the 11th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 321-331).

Harman, D. (1993). Overview of the first TREC conference. *Proceedings of the 16th Annual ACM SIGIR Conference of Research and Development in Information Retrieval*. (pp. 36-47).

Harman, D. (1996). Overview of the fourth Text REtrieval Conference. *Proceedings of the 4th Text Retrieval Conference*. (pp. 1-23).

Harman, D. (2002). TREC-2002 Novelty Track Report. *Text Retrieval Conference*.

Harter, S. P. (1992). Psychological relevance for information science. *Journal of the American Society for Information Science*, 43, 602-615.

Hearst, M. (1995). TileBars: Visualization of term distribution information in full text information access. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. (pp. 59-66).

Hemmje, M. (1995). LyberWorld: A 3D graphical user interface for fulltext retrieval. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (Conference Companion)*. (pp. 417-418).

Hersh, W. and Over, P. (2001). TREC-10 Interactive Track Report. *Text Retrieval Conference*.

Holscher, C. and Strube, G. (2000). Web search behaviour of internet experts and newbies. *Proceedings of the World Wide Web Conference*. (pp. 337-346).

Ikehara, C. S., Chin, D. N. and Crosby, M. E. (2003). A model for integrating an adaptive information filter utilizing biosensor data to assess cognitive load. *Proceedings of the 9th International Conference on User Modeling*. (pp. 208-212).

Ingwersen, P. (1982). Search procedures in the library analysed from the cognitive point of view. *Journal of Documentation*, 38, 165-191.

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

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities : Elements of a cognitive theory for information retrieval interaction. *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 101-110).

Iwayama, M. (2000). Relevance feedback with a small number of relevance judgements: incremental relevance feedback vs. document clustering. *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 10-16).

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

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

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

Joachims, T., Freitag, D. and Mitchell, T. (1997). WebWatcher: A tour guide for the world wide web. *Proceedings of the 16th Joint International Conference on Artificial Intelligence*. (pp. 770-775).

Jones, W., Bruce, H. and Dumais, S. (2001). Keeping found things found on the web. *Proceedings of the 10th Conference on Information and Knowledge Management*. (pp. 119-134).

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

Kelly, D. and Belkin, N. J. (2001). Reading time, scrolling and interaction: Exploring sources of user preferences for relevance feedback during interactive information retrieval. *Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 408-409).

Kelly, D. and Belkin, N. J. (2002). A user modelling system for personalized interaction and tailored retrieval in interactive IR. *Proceedings of the Annual Conference of the American Society for Information Science and Technology*. (pp. 316-325).

Kelly, D. and Belkin, N. J. (2004). Display time as implicit feedback: Understanding task effects. *Proceedings of the 27th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 377-384).

Kelly, D. and Cool, C. (2002). The effects of topic familiarity on information search behavior. *Proceedings of the 2nd ACM/IEEE Joint Conference on Digital Libraries*. (pp. 74-75).

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

Kelly, G. A. (1963) *A theory of personality: The psychology of social constructs*. New York: Norton.

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

Kirsh, D. (2000). A few thoughts on cognitive overload. *Intellectia*.

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

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R. and Reidl, J. (1997). GroupLens: Applying collaborative filtering of news. *Communications of the ACM*, 40 (3), 77-87.

Kuhlthau, C. (1988). Developing a model of the library search process: Cognitive and affective aspects. *Retrieval Quarterly*, 28 (2), 232-242.

Kuhlthau, C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42 (5), 361-371.

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

Kuhlthau, C. (1999). The role of experience in the information search process of an early career information worker: Perceptions of uncertainty, complexity, construction and sources. *Journal of the American Society for Information Science*, 50 (5), 399-412.

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

Kupiec, J., Pedersen, J. and Chen, F. (1995). A trainable document summarizer. *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 68-73).

Kupperman, M. (1960). On comparing two observed frequency counts. *Applied Statistics*, 9, 37-42.

Lam, W., Mukhopadhyay, S., Mostafa, J. and Palakal, M. (1996). Detection of shifts in user interests for personalised information filtering. *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 317-325).

Lam-Adesina, A. M. and Jones, G. J. F. (2001). Applying summarization techniques for term selection in relevance feedback. *Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 1-9).

Landow, G. P. (1987). Relationally encoded links and the rhetoric of hypertext. *Proceedings of HyperText '87*. (pp. 331-338).

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

Lashkari, T., Metral, M. and Maes, P. (1994). Collaborative interface agents. *Proceedings of the American Association for Artificial Intelligence*. (pp. 444-449).

Lee, L. (1999). Measures of distributional similarity. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. (pp. 25-32).

Lieberman, H. (1995). Letizia: An agent that assists web browsing. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. (pp. 475-480).

Lieberman, H., Fry, C. and Weitzman, L. (2001). Exploring the web with reconnaissance agents. *Communications of the ACM*, 44 (7), 69-75.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Processing and Development.*, 2 (2), 159-165.

Mackay, D. M. (1960). What makes the question? *The Listener*, 62, 789-790.

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, 37 (7), 30-40.

Magennis, M. and van Rijsbergen, C. J. (1998). The potential and actual effectiveness of interactive query expansion. *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 324-332).

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

Maglio, P. P., Barrett, R., Campbell, C. S. and Selker, T. (2000). SUITOR: An attentive information system. *Proceedings of the Annual Conference on Intelligent User Interfaces*. (pp. 169-176).

Marchionini, G. (1995) *Information seeking in electronic environments*. Cambridge: Cambridge University Press.

Marchionini, G. and Shneiderman, B. (1998). Finding Facts vs. browsing knowledge in hypertext systems. *IEEE Computer*, 70-80.

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

Maron, M. E. and Kuhns, J. L. (1960). On relevance, probabilistic indexing and retrieval. *Journal of the ACM*, 7, 216-244.

McKeown, K., Robin, J. and Kukich, K. (1995). Generating concise natural language summaries. *Information Processing and Management*, 31 (5), 735-751.

Meddis, R. (1984) *Statistics using ranks: A unified approach*. Oxford: Basil Blackwell.

Mellon, C. A. (1986). Library anxiety: A grounded theory and its development. *College and Research Libraries*, 47, 160-165.

Miller, B. N., Riedl, J. T. and Konstan, J. A. (2003). GroupLens for Usenet: Experienced in applying collaborative filtering to a social interaction system. (Eds, Lueg, C. and Fisher, D.), *From Usenet to CoWebs: Interacting with Social Information Spaces* (pp. 206-231). London: Springer Press.

Mitchell, T., Caruana, R., Freitag, D., McDermott, J. and Zabowski, D. (1994). Experience with a learning personal assistant. *Communications of the ACM*, 37 (7), 81-91.

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

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers*, 10 (3), 305-322.

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

Mostafa, J., Mukhopadhyay, S. and Palakal, M. (2003). Simulation studies of different dimensions of users' interests and their impact on user modelling and information filtering. *Information Retrieval*, 6, 199-223.

Muller, A. and Thiel, U. (1994). Query expansion in an abductive information retrieval system. *Proceedings of RIAO Conference on Content-Based Multimedia Access*. (pp. 461-480).

Muramatsu, J. and Pratt, W. (2001). Transparent queries: Investigating users' mental models of search engines. *Proceedings of the 24th Annual ACM Conference on Research and Development in Information Retrieval*. (pp. 217-224).

Nichols, D. M. (1997). Implicit ratings and filtering. *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*. (pp. 31-36).

Oard, D. and Kim, J. (2001). Modeling information content using observable behaviors. *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*. (pp. 38-45).

Oddy, R. N. (1977). Information retrieval through man-machine dialogue. *Journal of Documentation*, 33 (1), 1-14.

Osdin, R., Ounis, I. and White, R. W. (2002). Using hierarchical clustering and summarisation approaches for web retrieval: Glasgow at the TREC 2002 interactive track. *Proceedings of the TREC 2002 Interactive Track*.

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

Paice, C. D. (1981). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. (Eds, Oddy, R. N., Robertson, S. E., van Rijsbergen, C. J. and Williams, P. W.), *Information Retrieval Research* (pp. 172-191). Butterworths.

Paice, C. D. (1990). Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management*, 26 (1), 171-186.

Pao, M. L. (1993). Term and citation retrieval. *Information Processing and Management*, 29 (1), 95-112.

Park, T. K. (1993). The nature of relevance in information retrieval: An empirical study. *Library Quarterly*, 63, 318-351.

Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: Networks of plausible reasoning*. San Mateo, CA: Morgan Kaufmann.

Picard, R. (1997) *Affective computing*. Cambridge, MA: MIT Press.

Picard, R. W. and Klein, J. (2002). Computers that recognise and respond to user emotion: Theoretical and practical implications. *Interacting with Computers*, 14 (2), 141-169.

Pirolli, P. and Card, S. (1995). Information foraging in information access environments. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. (pp. 51-58).

Pors, N. O. (2000). Information retrieval, experimental models and statistical Analysis. *Journal of Documentation*, 56 (1), 55-70.

Radev, D. R. and McKeown, K. R. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24 (3), 469-500.

Rath, G. J., Resnick, A. and Savage, R. (1961). The formation of abstracts by the selection of sentences: Part I - sentence selection by men and machines. *American Documentation*, 12 (2), 139-141.

Rees, A. M. (1967). Evaluation of information systems and services. *Annual Review of Information Science and Technology*, 2, 63-86.

Rich, E. (1983). Users are individuals: Individualizing user models. *International Journal of Human-Computer Studies*, 51, 323-338.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation*, 46 (4), 359-364.

Robertson, S. E. and Hancock-Beaulieu, M. M. (1992). On the evaluation of interactive IR systems. *Information Processing and Management*, 28 (4), 457-466.

Robins, D. (1997). Shifts of focus in information retrieval interaction. *Proceedings of the 65th Annual Meeting of the American Society for Information Science*, 123-134.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. (Ed, Salton, G.), *The SMART retrieval system - experiments in automatic document processing* (pp. 313-323).

Rodden, K. (1998). About 23 million documents match your query... *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (Doctoral Consortium).* (pp. 64-65).

Rush, J. E., Salvador, R. and Zamora, A. (1971). Automatic abstracting and indexing (II) Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science*, 22 (4), 220-274.

Ruthven, I. (2001). *Abduction, explanation and relevance feedback*. Unpublished doctoral dissertation, University of Glasgow, Glasgow, UK.

Ruthven, I. (2002). On the use of explanations as a mediating device for relevance feedback. *Proceedings of the 6th European Conference on Digital Libraries*. (pp. 338-345).

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

Ruthven, I., Lalmas, M. and Van Rijsbergen, C. J. (2002a). Combining and selecting characteristics of information use. *Journal of the American Society for Information Science and Technology*, 53 (5), 378-396.

Ruthven, I., Lalmas, M. and Van Rijsbergen, C. J. (2002b). Ranking expansion terms using partial and ostensive relevance. *Proceedings of the 4th International Conference on Conceptions of Library and Information Science*. (pp. 199-219).

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

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

Salton, G., Allan, J. and Buckley, C. (1993). Approaches to passage retrieval in full text retrieval systems. *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 49-58).

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

Salton, G., Singhal, A., Mitra, M. and Buckley, C. (1997). Automatic text structuring and summarisation. *Information Processing and Management.*, 33 (2), 193-207.

Saracevic, T. (1975). Relevance: A review of and a framework for thinking on the notion of information science. *Journal of the American Society for Information Science*, 26 (6), 321-343.

Saracevic, T. (1996). Relevance reconsidered '96. *Proceedings of the 2nd International Conference on Conceptions of Library and Information Science*. (pp. 201-218).

Saracevic, T., Kantor, P. B., Chamis, A. Y. and Trivison, D. (1988). A study on information seeking and retrieving: Part I - Background and methodology. *Journal of the American Society for Information Science*, 39 (3), 161-176.

Schamber, L. (1991). Users' criteria for evaluation in a multimedia environment. *Proceedings of the 59th Annual Meeting of the American Society for Information Science*. (pp. 3-9).

Schamber, L., Eisenberg, M. and Nilan, M. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management*, 26 (6), 755-776.

Seo, Y. W. and Yang, B. T. (2000). Learning user's preferences by analysing web browsing behaviors. *Proceedings of the 4th ACM International Conference on Autonomous Agents*. (pp. 381-387).

Sheth, B. and Maes, P. (1993). Evolving agents for personalized information filtering. *Proceedings of the IEEE Conference on Artificial Intelligence for Applications.* (pp. 345-352).

Shneiderman, B. (1998) *Designing the user interface: Strategies for effective human-computer interaction*. Boston: Addison-Wesley.

Shneiderman, B., Byrd, D. and Croft, W. B. (1997). Clarifying search: A user-Interface framework for text searches. *D-Lib Magazine*.

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

Silverstein, C., Henzinger, M., Hannes, M. and Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, 33 (1), 6-12.

Smeaton, A. (1990). Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal*, 35, 268-278.

Smithson, S. (1990). The evaluation of information retrieval systems: A case study approach. (Ed, Jones, K. P.), *Infomatics 10 - Prospects for intelligent retrieval* (pp. 75-79). London: Aslib.

Spärck-Jones, K. (1981). Retrieval system tests 1958-1978. (Ed, Spärck-Jones, K.), *Information retrieval experiment* (pp. 213-255). London: Butterworths.

Spärck-Jones, K. and Endres-Niggermeyer, B. (1995). Automatic summarizing. *Information Processing and Management*, 31 (5), 625-630.

Spink, A. (1996). Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science*, 47 (8), 603-609.

Spink, A., Goodrum, A., Robins, D. and Wu, M. M. (1996). Search intermediaries elicitations during mediated online searching. *Proceedings of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 120-127).

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

Spink, A., Jansen, B. J., Wolfram, D. and Saracevic, T. (2002). From E-Sex to E-Commerce: Web search changes. *IEEE Computer*, 107-109.

Spink, A. and Losee, R. M. (1996). Feedback in information retrieval. *Annual Review of Information Science and Technology*, 31, 33-78.

Spink, A. and Saracevic, T. (1997). Interactive information retrieval: Sources and effectiveness of search terms during mediated online searching. *Journal of the American Society for Information Science*, 48 (8), 741-761.

Spoerri, A. (1993). InfoCrystal. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (Conference companion)*. (pp. 11-12).

Strzalkowski, T., Wang, J. and Wise, B. (1998). Summarization-based query expansion in information retrieval. *Proceedings of the 17th International Conference on Computational Linguistics*. (pp. 1-21).

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing and Management.*, 28 (4), 503-516.

Swanson, D. R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *Library Quarterly*, 56, 389-398.

Tague, J. and Schultz, R. (1988). Some measures and procedures for evaluation of the user interface in an information retrieval system. *Proceedings of the 11th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 371-385).

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, 28 (4), 467-490.

Tait, J. I. (1985). Generating summaries using a script based language analyser. (Eds, Steels, L. and Campbell, J. A.), *Progress in artificial intelligence* (pp. 312-318). Chichester: Ellis Horwood.

Tang, R. and Solomon, P. (1998). Towards an understanding of the dynamics of relevance judgements: An analysis of one person's search behaviour. *Information Processing and Management*, 43 (2/3), 237-256.

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

Teufel, S. and Moens, M. (1997). Sentence extraction as a classification task. *Proceedings of the ACL/EACL Summarization Workshop*. (pp. 58-65).

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

Tombros, A., Jose, J. M. and Ruthven, I. (2003a). Clustering top-ranking sentences for information access. *Proceedings of the 7th European Conference on Digital Libraries*. (pp. 523-528).

Tombros, A., Jose, J. M., Ruthven, I. and White, R. W. (2003b). Clustering the information space using top-ranking sentences: A study of user interaction. *Proceedings of the 9th INTERACT Conference*. (pp. 928-931).

Tombros, A., Ruthven, I. and Jose, J. M. (2003c). How users assess web pages for information seeking. *Journal of the American Society for Information Science and Technology*.

Tombros, A. and Sanderson, M. (1998). Advantages of query-biased summarisation in information retrieval. *Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 2-10).

Turtle, H. (1994). Natural language vs. boolean query evaluation: A comparison of retrieval performance. *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 212-220).

Turtle, H. and Croft, W. B. (1992). A comparison of text retrieval methods. *The Computer Journal*, 35 (3), 279-289.

Vakkari, P. (1998). Growth of theories on information seeking: An analysis of growth of a theoretical research program on the relation between task complexity and information seeking. *Information Processing and Management*, 34 (2/3), 361-382.

Vakkari, P. (1999). Task complexity, problem structure and information actions: Integrating studies in on information seeking and retrieval. *Information Processing and Management*, 35, 819-837.

Vakkari, P. (2001). A theory of task-based information retrieval. *Journal of Documentation*, 57 (1), 44-60.

Vakkari, P. (2002). Subject knowledge, source of terms, and term selection in query expansion: An analytical study. *Proceedings of the 24th Annual European Colloquium on Information Retrieval Research*. (pp. 110-123).

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology*, 37, 413-464.

Vakkari, P. and Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation*, 56 (5), 540-562.

Van Rijsbergen, C. J. (1986). A new theoretical framework for information retrieval. *Proceedings of the 10th International ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 194-200).

Van Rijsbergen, C. J. (1992). Probabilistic retrieval revisited. *The Computer Journal*, 35 (3), 291-298.

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

Voorhees, E. H. and Harman, D. (2000). Overview of the sixth text retrieval conference (TREC-6). *Information Processing and Management*, 36 (1), 3-35.

Wehrle, T. and Kaiser, S. (2000). Emotion and facial expression. *Affective Interactions: Towards a new generation of computer interfaces*. (pp. 49-63).

Weis, R. L. and Katter, R. V. (1967) *Multidimensional scaling of documents and surrogates.* Santa Monica: Systems Development Corporation.

White, R. W. (2004). A visualisation technique to communicate implicit feedback decisions. *Proceedings of the 26th European Conference on Information Retrieval (Vol. 2)*. (pp. 23-24).

White, R. W. and Jose, J. M. (2004). A study of topic similarity measures. *Proceedings of the 27th Annual ACM Conference on Research and Development in Information Retrieval*. (pp. 520-521).

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

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

White, R. W., Jose, J. M. and Ruthven, I. (2004a). An implicit feedback approach for interactive information retrieval. *Information Processing and Management*, in press.

White, R. W., Jose, J. M., van Rijsbergen, C. J. and Ruthven, I. (2004b). A simulated study of implicit feedback models. *Proceedings of the 26th Annual European Conference on Information Retrieval*. (pp. 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. *Proceedings of the 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 57-64).

White, R. W., Ruthven, I. and Jose, J. M. (2002b). The use of implicit evidence for relevance feedback in web retrieval. *Proceedings of 24th BCS-IRSG European Colloquium on Information Retrieval Research*. (pp. 93-109).

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval*, 9, 457-469.

Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation*, 36 (7), 3-15.

Woodruff, A., Faulring, A., Rosenholtz, R., Morrison, J. and Pirolli, P. (2001). Using thumbnails to search the web. *Proceedings of the ACM SIGCHI Conference on Human Factors In Computing Systems*. (pp. 198-205).

Wurman, R. S. (1989) *Information anxiety*. New York: Doubleday Books.

Zamir, O. and Etzioni, O. (1999). Grouper: A dynamic clustering interface to Web search results. *Proceedings of the 8th International Conference on World Wide Web*. (pp. 1361 - 1374).

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

Published Work

Journals

1. White, R.W., Jose, J.M. and Ruthven, I. (2004). An implicit feedback approach for interactive information retrieval. *Information Processing and Management*, in press.
2. White, R.W., Jose, J.M. and Ruthven, I. (2004). Using Top-Ranking Sentences to facilitate effective information access. *Journal of the American Society for Information Science and Technology*, in press.
3. White, R.W., Jose, J.M. and Ruthven, I. (2003). The influencing effects of query-biased summarisation in web searching. *Information Processing and Management*, 39 (5), 707-733.

Conference Papers

4. White, R.W., Jose, J. M., van Rijsbergen, C.J. and Ruthven, I. (2004). A simulated study of implicit feedback models. *Proceedings of the 26th Annual European Conference on Information Retrieval*, (pp. 311-326) (**Best Student Paper**).
5. White, R.W., Jose, J. M. and Ruthven, I. (2003). A granular approach to web search result presentation. *Proceedings of the 9th IFIP TC13 Conference on Human Computer Interaction*, (pp. 213-220) (**The Brian Shackel Award for Best Paper**).
6. Tombros, A., Jose, J. M., Ruthven, I. and White, R. W. (2003). Clustering the information space using Top-Ranking Sentences: A study of user interaction. *Proceedings of the 9th IFIP TC13 Conference on Human Computer Interaction*. (pp. 928-931).
7. Osdin, R., Ounis, I. and White, R. W. (2002). Using hierarchical clustering and summarisation approaches for web retrieval: Glasgow at the TREC 2002 interactive track. *Proceedings of the TREC 2002 Interactive Track*.
8. White, R. W., Ruthven, I. and Jose, J. M. (2002). Finding relevant web documents using top ranking sentences: An evaluation of two alternative schemes. *Proceedings of the 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 57-64).
9. White, R. W., Ruthven, I. and Jose, J. M. (2002). The use of implicit evidence for relevance feedback in web retrieval. *Proceedings of 24th BCS-IRSG European Colloquium on Information Retrieval Research*. (pp. 93-109).
10. White, R.W., Jose, J.M. and Ruthven, I. (2002). Implicit contextual modelling for information seeking. *Proceedings of the Glasgow Context Group 1st Colloquium: Building Bridges: Interdisciplinary Context-Sensitive Computing*.
11. White, R.W., Jose, J. and Ruthven, I. (2001). Comparing implicit and explicit feedback techniques for web retrieval : TREC-10 interactive track report. *Proceedings of the 10th Text Retrieval Conference*. (pp. 534-538).

Workshops

12. Haggerty, A., White, R.W. and Jose, J.M. (2003). NewsFlash: Adaptive TV news delivery on the Web. *Proceedings of the 1st International Workshop on Adaptive Multimedia Retrieval*. (pp. 72-86).
13. White, R.W., Ruthven, I. and Jose, J.M. (2001). Web document summarisation: A task-oriented evaluation. *Proceedings of the 1st International Workshop on Digital Libraries*. (pp. 951-955).

Conference Posters

14. White, R. W. and Jose, J. M. (2004). A study of topic similarity measures. *Proceedings of the 27th Annual ACM Conference on Research and Development in Information Retrieval*. (pp. 520-521).
15. White, R. W. (2004). A visualisation technique to communicate implicit feedback decisions. *Proceedings of the 26th European Conference on Information Retrieval (Vol. 2)*. (pp. 23-24).
16. White, R.W., Jose, J.M. and Ruthven, I. (2003). Adapting to evolving needs: Evaluating a behaviour-based search interface. *Proceedings of the 17th Annual Human-Computer Interaction Conference*. (pp. 125-128).
17. White, R.W., Jose, J.M. and Ruthven, I. (2003). Using Top-Ranking Sentences for web search result presentation. *Proceedings of the 12th International World Wide Web Conference*.

18. White, R.W., Jose, J.M. and Ruthven, I. (2003). An approach for implicitly detecting information needs. *Proceedings of the 12th Annual ACM CIKM Conference on Information and Knowledge Management.* (pp. 504-507).
19. White, R.W., Jose, J.M. and Ruthven, I. (2001). Query-biased web page summarisation: A task-oriented evaluation. *Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval.* (pp. 412-413).

Demonstrations

20. White, R.W. and Jose, J.M. (2004). An implicit system for predicting interests. *Proceedings of the 27th Annual ACM SIGIR Conference on Research and Development in Information Retrieval.* (p. 590).
21. White, R.W., Jose, J.M. and Ruthven, I. (2002). A system using implicit feedback and top ranking sentences to help users find relevant web documents. *Proceedings of the 25th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 446).

Appendices

Appendix A

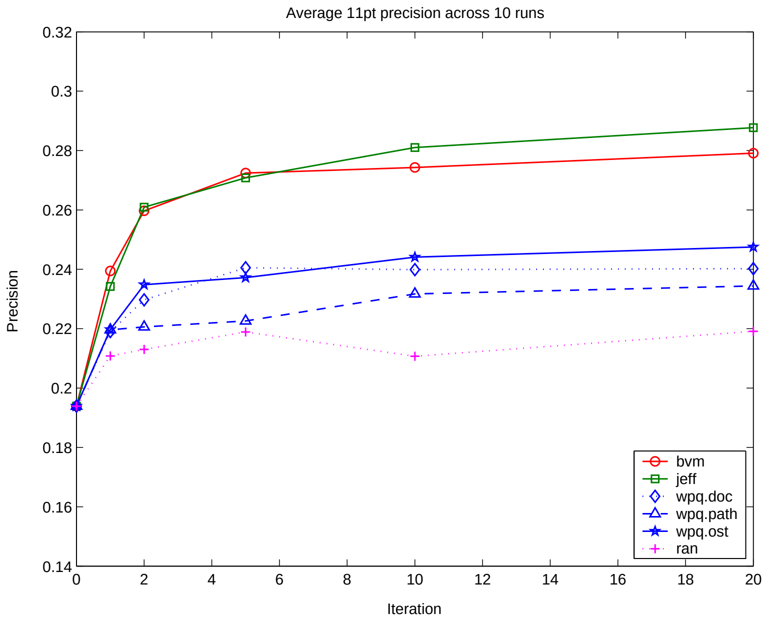


Figure A.1. Average 11-point precision across 10 runs for 10% wandering.

Table A.1. Percentage change in precision per iteration for a wandering level of 10%. 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	**19.0**	–	25.3	+ 7.8	**28.8**	+ **4.7**	29.3	+ 0.7	30.5	+ 1.7
jeff	17.2	–	**25.7**	+ **10.2**	28.4	+ 3.7	**31.0**	+ 3.6	**32.6**	+ 2.3
wpq.doc	11.5	–	15.6	+ 4.7	19.4	+ 4.5	19.4	– 0.3	19.3	+ 0.1
wpq.path	11.7	–	12.1	+ 0.5	12.9	+ 0.9	16.3	+ **3.9**	17.3	+ 1.2
wpq.ost	11.7	–	17.4	+ 6.4	18.3	+ 1.0	20.6	+ 2.8	21.7	+ 1.4
ran	8.0	–	9.0	+ 1.0	11.4	+ 2.7	8.0	– 3.9	11.5	+ **3.8**

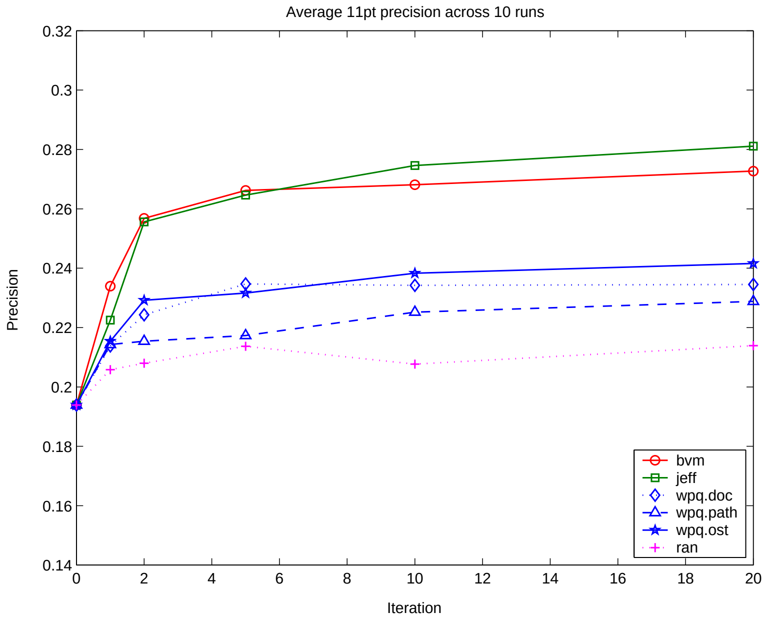


Figure A.2. Average 11-point precision across 10 runs for 20% wandering.

Table A.2. Percentage change in precision per iteration for a wandering level of 20%. 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	**17.1**	–	**24.5**	+ 8.9	**27.2**	+ 3.5	27.7	+ 0.7	28.9	+ 1.7
jeff	12.9	–	24.1	+ **12.9**	26.7	+ 3.4	**29.4**	+ **3.6**	**31.0**	+ 2.3
wpq.doc	9.3	–	13.6	+ 4.7	17.4	+ **4.4**	17.2	– 0.2	17.3	+ 0.1
wpq.path	9.5	–	10.0	+ 0.5	10.8	+ 0.9	13.9	+ 3.5	15.3	+ 1.6
wpq.ost	10.0	–	15.4	+ 6.0	16.3	+ 1.0	18.6	+ 2.8	19.7	+ 1.4
ran	5.8	–	6.8	+ 1.1	9.3	+ 2.7	6.6	– 2.9	9.4	+ **2.9**

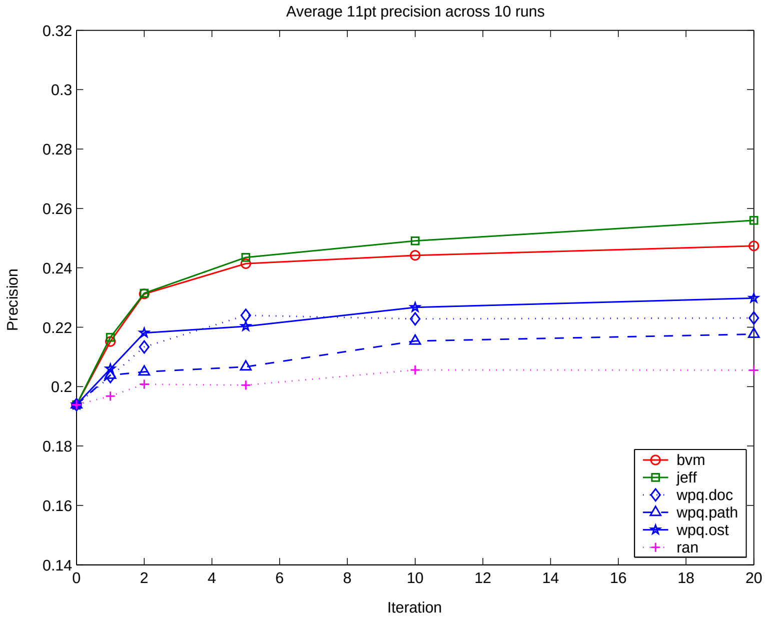


Figure A.3. Average 11-point precision across 10 runs for 30% wandering.

Table A.3. Percentage change in precision per iteration for a wandering level of 30%. 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	9.9	–	16.1	+ **6.9**	19.7	+ 4.2	20.6	+ 1.1	21.6	+ 1.3
jeff	**10.5**	–	**16.2**	+ 6.4	**20.4**	+ **5.0**	**22.2**	+ 2.2	**24.2**	+ 2.7
wpq.doc	4.7	–	9.1	+ 4.7	13.4	+ 4.8	13.0	– 0.5	13.1	+ 0.1
wpq.path	4.9	–	5.4	+ 0.5	6.2	+ 0.8	10.0	+ **4.0**	10.9	+ 1.1
wpq.ost	5.9	–	11.1	+ 5.6	12.0	+ 1.0	14.5	+ 2.8	15.6	+ 1.4
ran	1.5	–	3.4	+ 2.0	5.5	+ 2.2	0.9	– 4.9	5.6	+ **4.8**

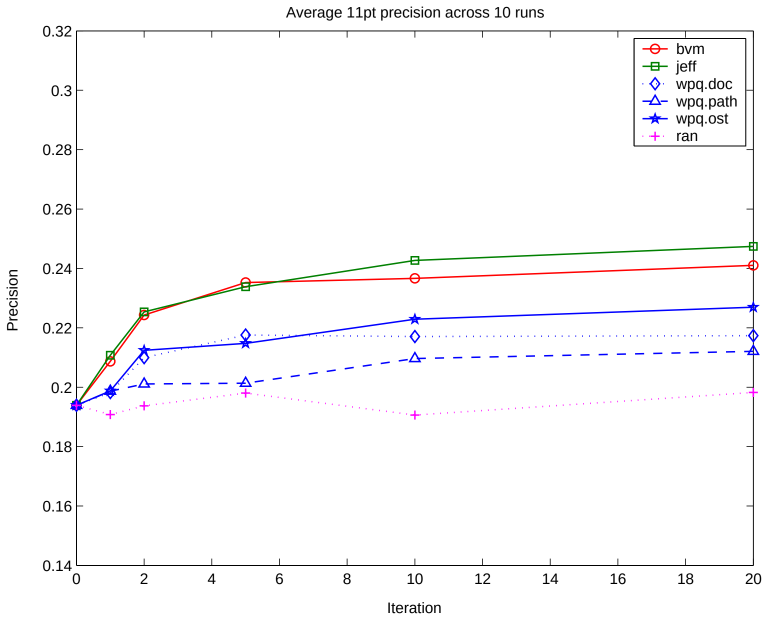


Figure A.4. Average 11-point precision across 10 runs for 40% wandering.

Table A.4. Percentage change in precision per iteration for a wandering level of 40%. 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	7.1	–	13.6	+ **6.9**	**17.6**	+ **4.6**	18.1	+ 0.6	19.5	+ 1.8
jeff	**8.0**	–	**14.0**	+ 6.5	17.1	+ 3.6	**20.1**	+ 3.6	**21.6**	+ 1.9
wpq.doc	2.2	–	7.6	+ 5.9	10.9	+ 3.5	10.7	– 0.2	10.8	+ 0.1
wpq.path	2.4	–	3.6	+ 1.2	3.7	+ 0.1	7.5	+ **3.9**	8.6	+ 1.1
wpq.ost	2.4	–	8.7	+ 6.4	9.7	+ 1.1	13.0	+ 3.6	14.5	+ 1.8
ran	– 1.6	–	– 0.1	+ 1.5	2.1	+ 2.2	– 1.7	– 3.9	2.2	+ **3.9**

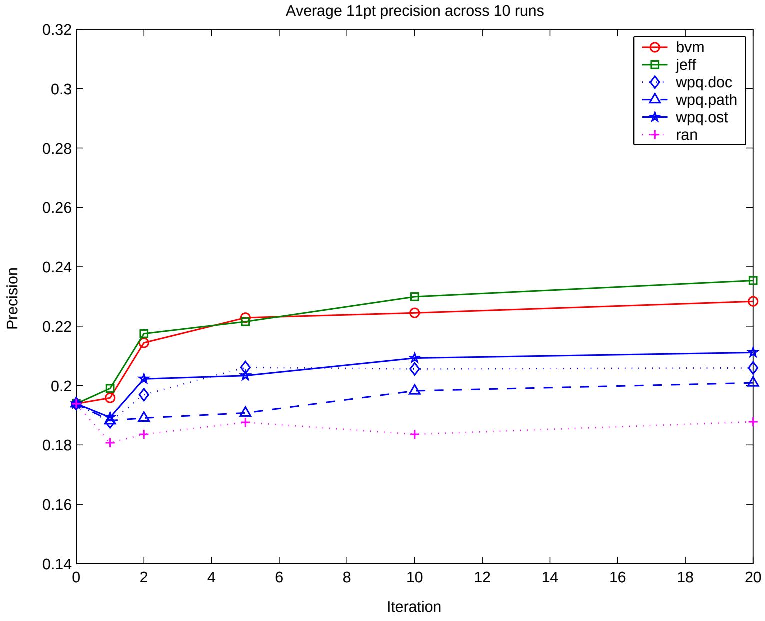


Figure A.5. Average 11-point precision across 10 runs for 50% wandering.

Table A.5. Percentage change in precision per iteration for a wandering level of 50%. 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	1.0	–	9.6	+ **8.7**	**13.0**	+ 3.8	13.6	+ 0.7	15.1	+ 1.7
jeff	**2.6**	–	**10.8**	+ 8.5	12.5	+ 1.8	**15.7**	+ 3.6	**17.6**	+ **2.3**
wpq.doc	– 3.3	–	1.5	+ 4.7	5.9	+ **4.5**	5.7	– 0.2	5.8	+ 0.2
wpq.path	– 3.0	–	– 2.5	+ 0.5	– 1.6	+ 0.9	2.2	+ **3.8**	3.5	+ 1.3
wpq.ost	– 2.4	–	4.1	+ 6.4	4.6	+ 0.5	7.3	+ 2.8	8.2	+ 0.9
ran	– 7.3	–	– 5.6	+ 1.6	– 3.4	+ 2.1	– 5.6	– 2.2	– 3.2	+ **2.3**

Appendix B

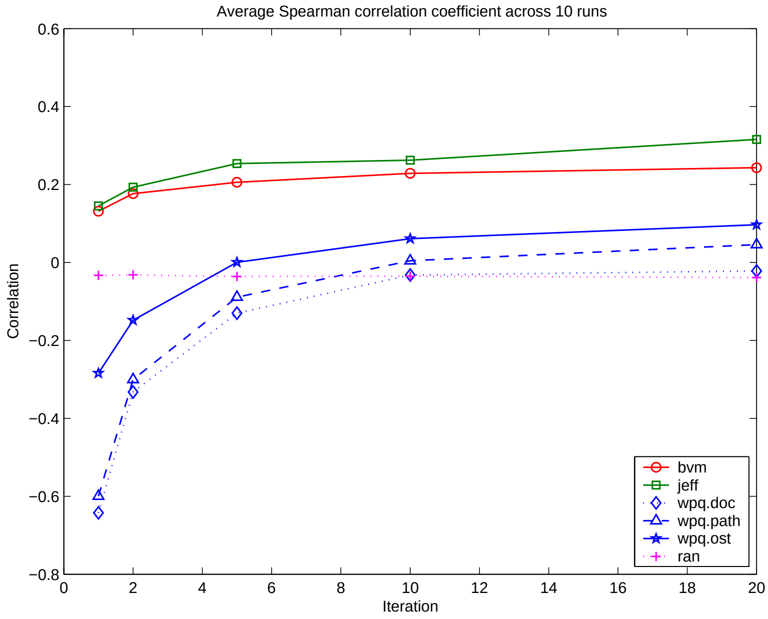


Figure B.1. Average Spearman correlation coefficient across 10 runs for 10% wandering.

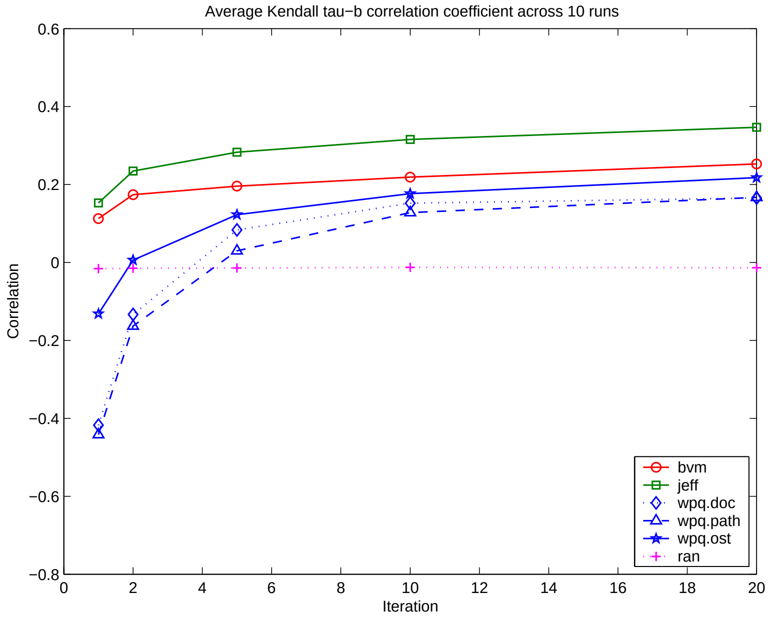


Figure B.2. Average Kendall correlation coefficient across 10 runs for 10% wandering.

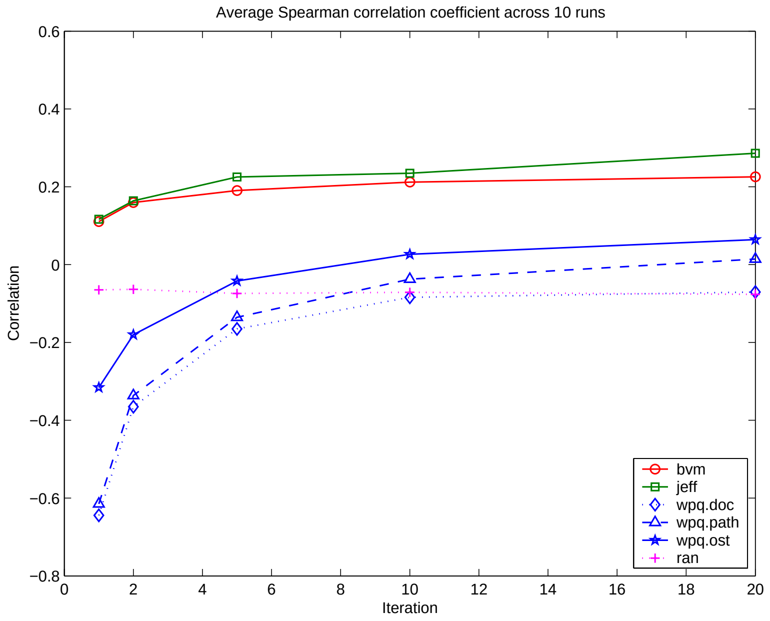


Figure B.3. Average Spearman correlation coefficient across 10 runs for 20% wandering.

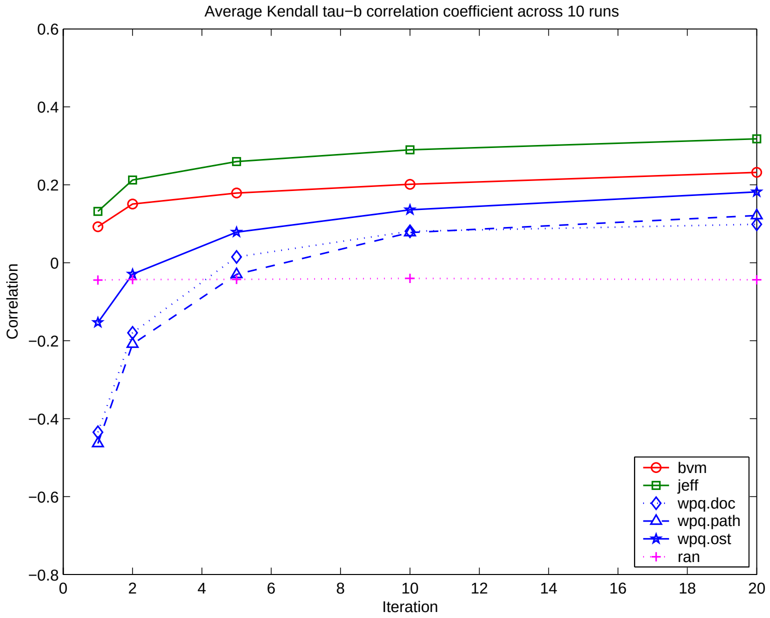


Figure B.4. Average Kendall correlation coefficient across 10 runs for 20% wandering.

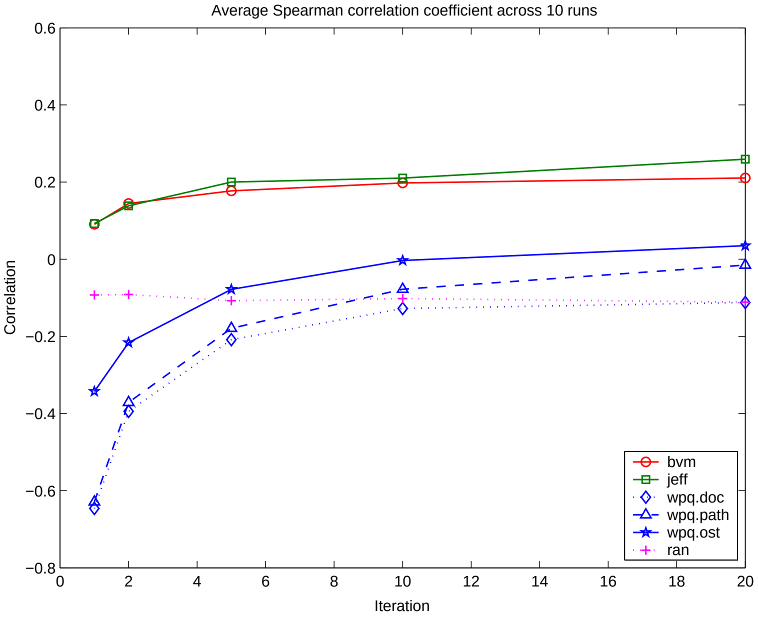


Figure B.5. Average Spearman correlation coefficient across 10 runs for 30% wandering.

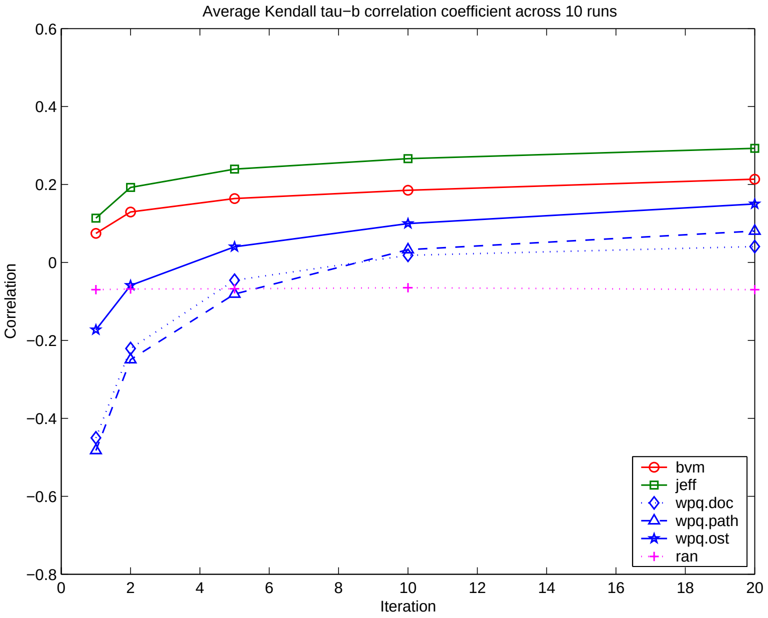


Figure B.6. Average Kendall correlation coefficient across 10 runs for 30% wandering.

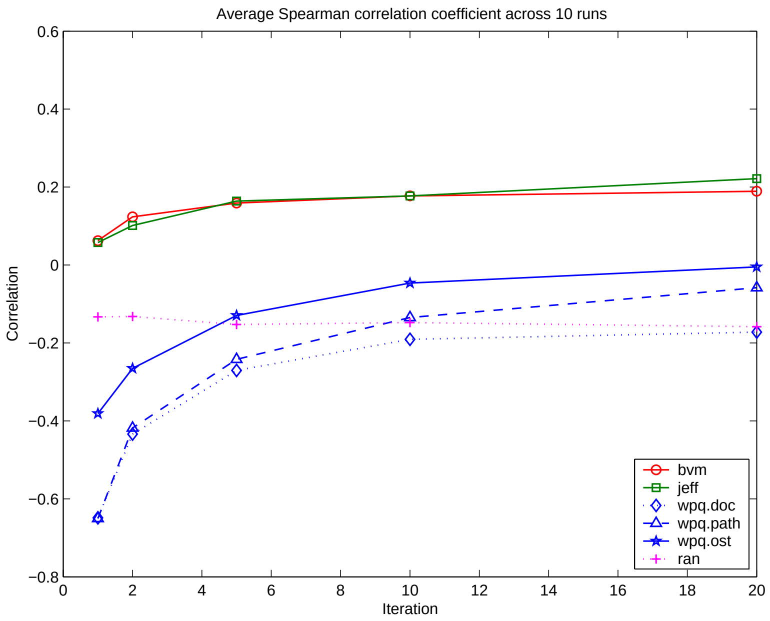


Figure B.7. Average Spearman correlation coefficient across 10 runs for 40% wandering.

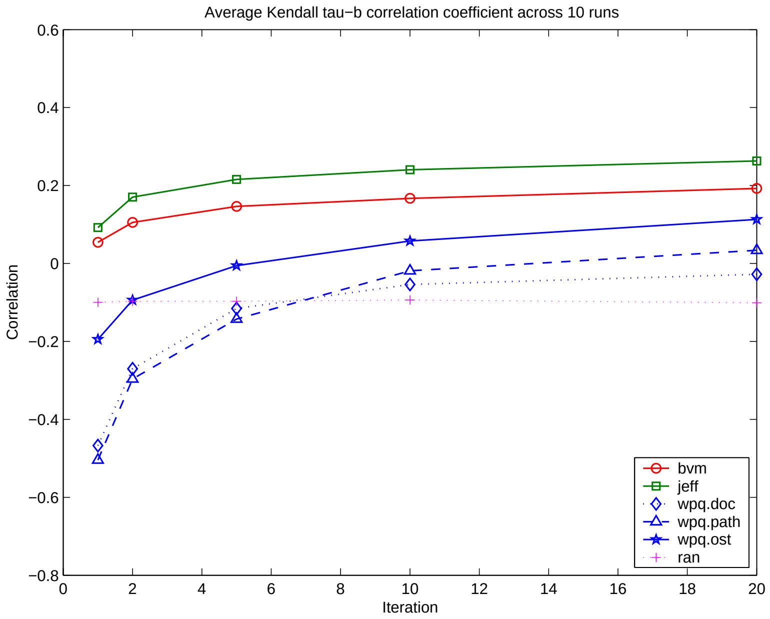


Figure B.8. Average Kendall correlation coefficient across 10 runs for 40% wandering.

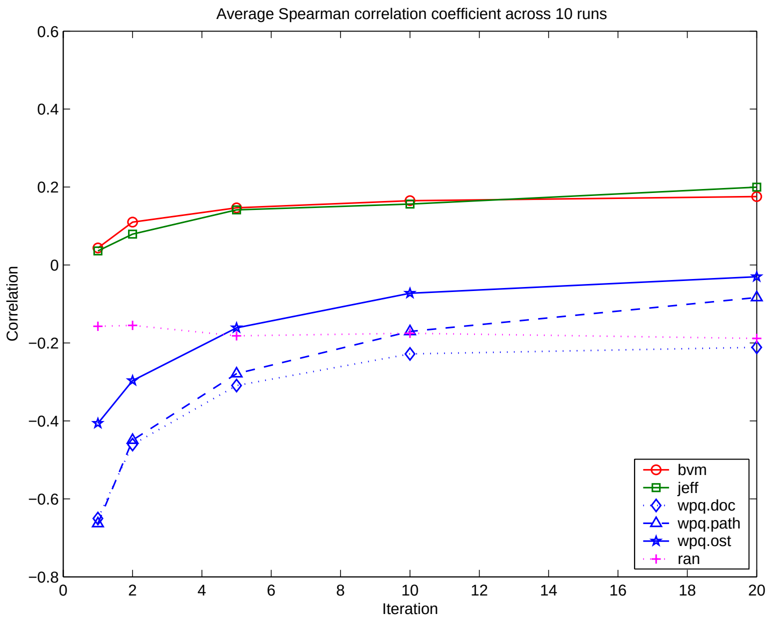


Figure B.9. Average Spearman correlation coefficient across 10 runs for 50% wandering.

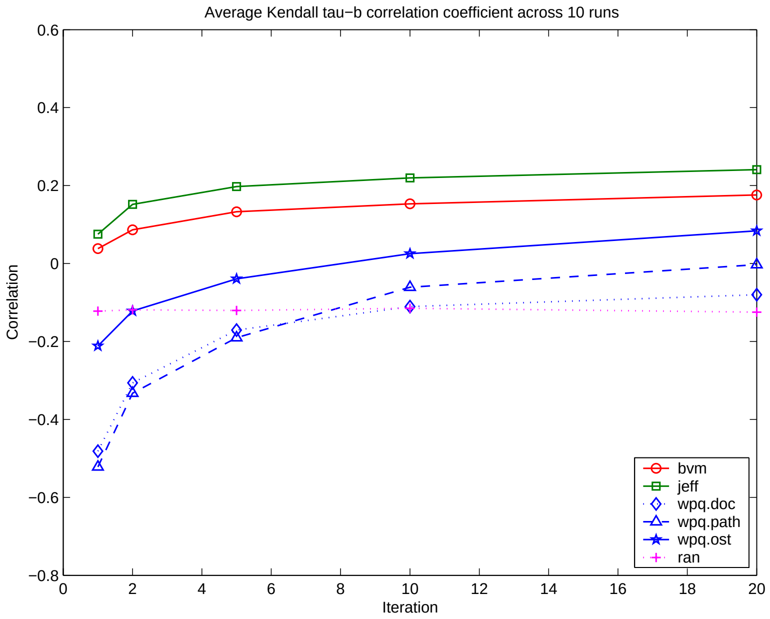


Figure B.10. Average Kendall correlation coefficient across 10 runs for 50% wandering.

Appendix C

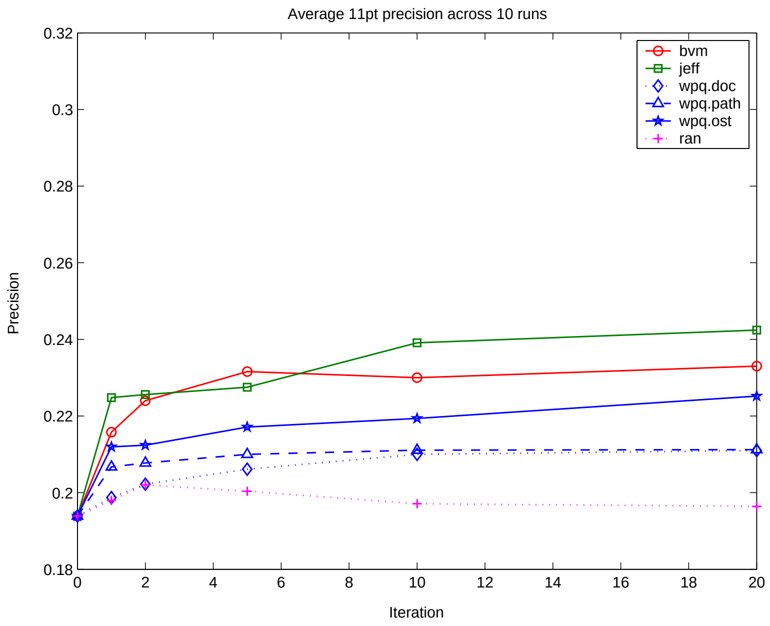


Figure C.1. Average 11-point precision across 10 runs for Scenario 5b.

Table C.1. Percentage change in precision per iteration for Scenario 5b. 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	10.1	–	13.4	+ **3.7**	**16.3**	+ **3.3**	15.7	– 0.7	16.8	+ 1.3
jeff	**13.7**	–	**14.1**	+ 0.4	14.8	+ 0.8	**18.9**	+ **4.9**	**20.0**	+ 1.4
wpq.doc	2.4	–	4.1	+ 1.7	5.9	+ 1.9	7.7	+ 1.9	8.1	+ 0.5
wpq.path	6.2	–	6.7	+ 0.5	7.7	+ 1.1	8.1	+ 0.5	8.2	+ 0.05
wpq.ost	8.5	–	8.7	+ 0.2	10.7	+ 2.2	11.6	+ 0.1	13.9	+ **2.6**
ran	2.1	–	4.0	+ 2.0	3.2	– 0.8	1.6	– 1.6	1.3	– 0.4

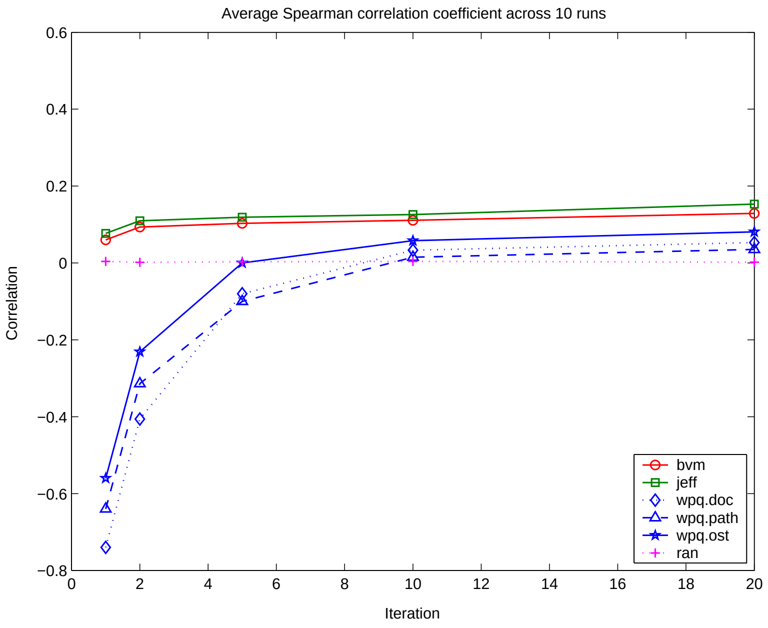


Figure C.2. Average Spearman correlation coefficient across 10 runs for Scenario 5b.

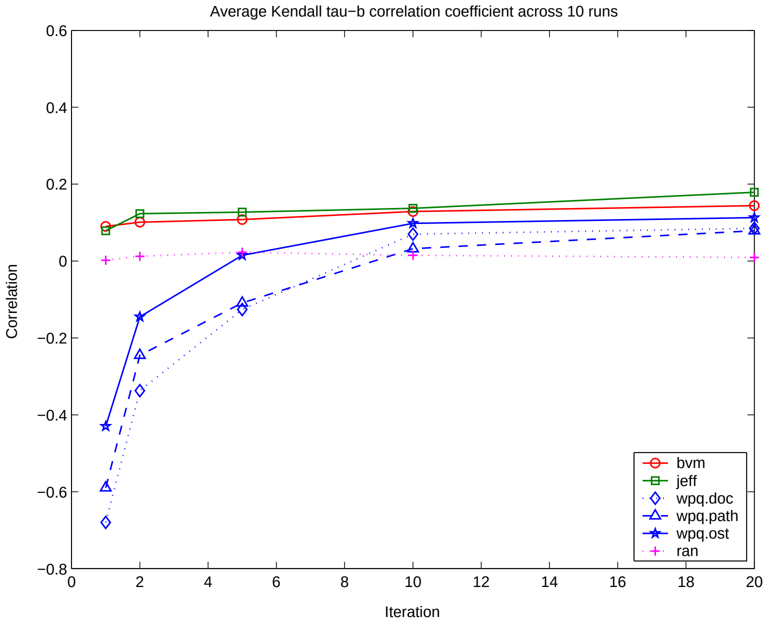


Figure C.3. Average Kendall correlation coefficient across 10 runs for Scenario 5b.

Appendix D

D.1 Introduction

The aim of this pilot test was to evaluate the interface components such as document representations and relevance paths and how well the heuristic-based framework identified information needs and tracked changes in formulations of them. The hypotheses tested were:

Hypothesis 1

> The terms selected by the heuristic-based framework identifies the information needs of the subject (i.e., term selection support).

Hypothesis 2a

> The heuristic-based framework tracks changes in the formulation of information needs.

Hypothesis 2b

> The heuristic-based framework makes search decisions that correspond closely with those of the subject.

These hypotheses tested the two components of the heuristic-based framework: the Binary Voting Model and the information need tracking component. Details now are given of the experimental subjects, the search tasks, the experimental methodology and the systems used.

D.2 Experimental Subjects

24 subjects were recruited. In a similar way to that already described in Chapter Four, recruitment was specifically aimed at targeting two groups of subjects: *inexperienced* and *experienced*. The experienced subjects were those who used computers and searched the Web on a regular basis. Inexperienced subjects were those who searched the Web, used computers and the Internet infrequently. On average per week, inexperienced subjects spent 3.1 hours online, and experienced subjects spent 34.9 hours online. Overall, subjects had an average age of 26 with a range of 38 years (youngest 16 years, oldest 54 years). 14 males and 10 females participated in the experiments. The classification between experienced and inexperienced subjects was made on the basis of the subjects’ responses to questions about their experience and their own opinion of their skill level.

D.3 Experimental Tasks

Each subject was asked to complete one search task from each of four categories, each containing two tasks. The categories were: *fact* search (e.g., finding a named person's current email address), *decision* search (e.g., choosing the *best* financial instrument), *background* search (e.g., finding information on dust allergies) and *search for a number of items* (e.g., finding contact details for a number of potential employers) (White *et al.*, 2003b). Each search task was placed within a simulated work task situation (Borlund, 2000b). This technique asserts that subjects should be given search scenarios that reflect real-life search situations and should allow the subject to make personal assessments on what constitutes relevant material. The different tasks engender realistic search behaviour and produce different types of simulated information needs within the range of verificative and conscious topical information needs (Ingwersen, 1992). The search tasks issued to subjects are included in Appendix E.

There were two tasks per category, each of a similar level of difficulty (verified by *a priori* pilot testing and questions in the post-task questionnaire) and subjects were asked to choose the task they would like to do. No other constraints were placed on their task selection. Subjects chose 51% of tasks because they were *interesting*, 21.8% of tasks because they felt they were *easy*, 19.8% because they were *familiar* with the topic area and 7.4% for *no reason*.

Offering subjects a choice of tasks allowed them to select tasks that interested them and were familiar. Subjects with topic experience are better equipped to make query modification decisions using that topic's terms and relevance assessments of that topic's documents (Kelly and Cool, 2002). Whilst the subject groups were homogeneous (i.e., inexperienced or experienced) no criteria other than search experience were used in the selection of candidates. Subject interests were potentially diverse and it was not possible to offer a single task in each task category that appealed to all subjects. Giving subjects a choice of tasks in each category increased the likelihood that the task they chose would interest them.

D.4 Experimental Methodology

In this pilot test subjects completed four search tasks, two tasks on each of the two systems: experimental and baseline. The presentation of tasks to subjects was held constant; each subject performed the search tasks in the same order, however the order of presentation of systems was rotated across subjects. The tasks had been used in previous experiments (White *et al.*, 2003b), where the impact of task bias was not significant. Subjects were given a maximum of 10 minutes to complete each task. Longitudinal evaluations such as those used in Vakkari (1999) and Kelly (2004) can be used to monitor search behaviours over periods of time and in operational environments. Since experimental and situational variables are difficult to control in longitudinal studies their usefulness for the comparative evaluation of retrieval systems is limited.

The subjects were given a short tutorial on the features of the two systems and a training task that allowed them to use both systems. Background data on aspects such as the subjects' experience and training in online searching was then captured using an 'Entry' questionnaire. After this, subjects were introduced to tasks and systems according to the experimental design. Subjects were instructed to attempt each task to the best of their ability and write their answer on a sheet provided. As it was felt that this may affect subject interaction, subjects were not told how the Binary Voting Model and information need tracking component operated. A search was regarded as successful if the subject felt they had succeeded in their performance of the task. This is closely related to real information seeking situations, where the goal of any retrieval system is to satisfy the searcher.

Once they had completed a search, the subject was asked to complete questionnaires regarding various aspects of the search. Semantic differentials, Likert scales and open-ended questions were used to collect these data. These methods for capturing subjective information have been effective in related work (Brajnik *et al.*, 1996). In addition, semi-structured interviews were conducted after each search and after the experiment as a whole. Background logging was used to record each subject interaction event (e.g., queries submitted, mouse clicks, etc.) with an associated time stamp.

D.5 Systems

Two systems were used in this pilot test: the experimental system and the baseline system. The systems used document representations and relevance paths in the same way as described in Chapter Five. The experimental system used subject interaction to infer interests, select appropriate terms to add to the query and make decisions about how to use the new query. The baseline system used the same interface components as the implicit system but differed in one key way; in the baseline system the searcher was solely responsible for adding new query terms and selecting what retrieval strategies were undertaken after these terms have been added. These options gave subjects increased control over the search but also increased responsibility for making decisions.

The baseline interface contains one additional component; a control panel that allowed subjects to modify their query and make search decisions. The nature of this baseline allowed me to evaluate how well the experimental system estimated information needs from the perspective of the subject. I tested whether the system chose terms and made search decisions that matched those selected by the subject and whether the subject felt the support offered was beneficial. Systems that use implicit feedback can be unpopular since they remove searcher burden but also searcher control (Kelly and Teevan, 2003). In this pilot test I acknowledged this, and compared the approach with a baseline where the subject has such control. In the next section the findings are discussed.

D.6 Discussion

From observations and informal post-search interviews, subjects appeared to use the relevance paths and find the information shown at the search interface of value in their search. This is important, as the success of the both systems – especially the experimental system – is dependent on the use of these interface features. This finding was also important for the design of the systems described later in this thesis as it demonstrated the potential of systems that structure and monitor searcher interaction in this way.

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

Subjects did not rate their own search terms as *always* useful. They acknowledge that they are not able to adequately conceptualise their information need, even when given the chance to refine the terms used to express it. However, as they view and process information, and their state of knowledge changes, they become more able to express these needs (Belkin, 1980). The Binary Voting Model through a process of reinforcement learning (i.e., being repeatedly shown indications of what constitutes relevance) learned progressively, training itself with searcher interaction to better identify what is relevant. The Binary Voting Model was used in this pilot test to test my initial ideas that were later formalised into the Jeffrey's Conditioning Model.

The Binary Voting Model chose terms to represent the information needs of the subject. I used the degree of term overlap between the terms chosen by the subject and those chosen by the system as a measure of how well the model approximated information needs. Across both subject groups terms chosen by the Binary Voting Model co-occurred with any subject terms on a high number of occasions.

All subjects were instructed before the experiment that the different search decisions provided varying degrees of interface support and will have an increasingly dramatic effect in recreating or restructuring the retrieved information. They were not told that the control related in any way to shifts, changes or developments in their information need. Subjects adapted well to the need tracking and seemed comfortable with making search decisions that led to different outcomes (i.e., re-searching the document collection or reorganising information already retrieved).

The form of implicit feedback tested in this pilot evaluation is at the extreme end of a spectrum of searcher support. Based on informal feedback received during and after the pilot test, the approach removed too much searcher control. Feedback systems that use implicit feedback techniques may be best used to make decisions in conjunction with, not in place of, the searcher. As in *interactive query expansion* (c.f. Koenemann and Belkin, 1996), a system implementing such technology would monitor interaction and present potentially useful terms at the interface. In this collaboration, the searcher – who is best equipped to make relevance decisions – would select potentially useful terms and add them to the search query. Subjects also suggested that the system could also recommend search decisions based on the predicted degree of information need change. The system would give the searcher control over whether the recommended strategy is then executed.

This test confirmed the value of the content-rich search interfaces and the effectiveness of the components to estimate information needs and information need change. A fuller description and analysis of Pilot Test 1 is presented in White *et al*. (2004a).

Appendix E

T1.Fact

Simulated work task situation: You have just finished reading a very interesting article from a popular journal in your area of research. It has been five years since the article was first published, but you note that the author is Jan-Jaap Ijdens from the Robert Gordon University, Aberdeen. You have a keen interest in what the article discusses and would like to send an electronic mail to the author. However, you contact the university and find that Dr Ijdens has moved, leaving no forwarding email address.

Task: Bearing in mind this context, your task is to find his current email address.

T2.Fact

Simulated work task situation: You have recently formed a quiz team with your friends at university and have decided to enter a national competition. As a precursor to being invited to participate, you must first answer a set of questions that will be sent off to the competition organisers to be marked. Only the top scoring teams will be invited to compete at the national finals. You are finding one of the questions on the identity of the first male winner of the New York Marathon difficult to answer as this is not your area of expertise. The only clue you are given is that it was first run in 1970.

Task: Bearing in mind this context, your task is to find the name of the first male winner of the New York Marathon.

T3.Decision

Simulated work task situation: This summer, during your vacation, you are planning to go on a touring trip of North America. You want information to help you plan your journey and there are many tourist attractions you would be interested in visiting. You have set aside 3 months for the trip and hope to see as much of the continent as you can. As you cannot drive, you will have to use public transport, but are unsure which type to take.

Task: Bearing in mind this context, your task is to decide on the *best* form of transportation between cities in North America that would suitable for you.

T4.Decision

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

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

T5.Background

Simulated work task situation: You have been asked, as part of your coursework for computing science or psychology this year, to write an essay on the Data Protection Act (computing) or 'Nature versus Nurture' (psychology). The essay should cite a number of sources, provide arguments for and against, and come to a conclusion incorporating your own views and opinions. You would like to gather information that could be useful for this task.

Task: Bearing in mind this context, your task is to find information that would be helpful for your essay, i.e., points for and against

T6.Background

Simulated work task situation: You are currently working as a research assistant at the University of Glasgow. Your laboratory is in an old building and one of your colleagues has developed a severe dust allergy that you believe is caused by his working environment. He is writing a letter to complain about the lack of cleanliness in your working environment and has asked you to help find information about dust allergies.

Task: Bearing in mind this context, your task is to find information about dust allergies in the workplace i.e., possible causes and cures.

T7.Number of Items

Simulated work task situation: You are entertaining a foreign exchange student who has expressed an interest in theatre and the arts. You are considering taking them to a local production of an Arthur Miller play. However, you are unfamiliar with his work and would like to find out more about the some of the plays he has written. You decide on three plays – 'The Crucible', 'Elegy for A Lady' and 'Death of a Salesman' – that you would be interested in finding more about.

Task: Bearing in mind this context, your task is to provide a one sentence description of the plot in each of the three plays.

T8.Number of Items

Simulated work task situation: After you graduate you will be looking for a job in industry. You would like to keep your options open as you are unsure of what you would like to do exactly. However, since your choice of subjects in subsequent years of your course will impact on your employment options, you feel that now is a good time to decide on a job that would suit you. Friends and family have advised you to contact employment agencies and companies working in career development.

Task: Bearing in mind this context, find five contact names and email addresses for such recruitment companies specialising in your preferred line of work.

Appendix F

In this Appendix I present the experimental documents from the experiment described in Part IV of this thesis. These include:

F.1. Information sheet, Consent form and Receipt of Payment

F.2. 'Entry', 'Search' and 'Exit' Questionnaires

F.3. Training Search Task, Search Tasks and Task Completion Sheet
(Task A: High Complexity, Task B: Moderate Complexity, Task C: Low Complexity)

Department: *Computing Science*
Subject Identification Number for this study:

UNIVERSITY
of
GLASGOW

INFORMATION SHEET

Title of Project:

Web Search Interface Investigation

Name of Researcher:

Ryen W. White

You are being invited to take part in a research study. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take the time to read the following information carefully. Ask me if there is anything that is not clear or if you would like more information.

The aim of this experiment is to investigate the relative effectiveness of three different Web search interfaces. We cannot determine the value of search systems unless we ask those people who are likely to be using them, which is why we need to run experiments like these. Please remember that it is the interfaces, not you, that are being evaluated. You were chosen, along with 47 others, because you work or study at the University of Glasgow.

It is up to you to decide whether or not to take part. If you decide to take part you will be given this information sheet to keep and asked to sign a consent form. If you decide not to take part you are free to withdraw at any time without giving a reason. You also have the right to withdraw retrospectively any consent given, and to require that any data gathered on you be destroyed. A decision not to participate will not affect your grades in any way.

The experiment will last between one-and-a-half and two hours and will you will receive a reward of £12 upon completion. You will be given a chance to learn how to use the three interfaces before we begin. At this time you will also be asked to complete an introductory questionnaire. You will perform three tasks, one with each interface, and complete a questionnaire about using each system. You will have 15 minutes for each task. The questionnaires will ask how you felt during each search. All of your interaction (e.g., mouse clicks, scrolling, key presses) will also be logged. You are encouraged to comment on each interface as you use it, all your comments will be recorded on audio cassette *or I will take notes if you so prefer*. You will have the option to review, edit, or erase the recording. Please ask questions if you need to and please let me know when you are finished each task. You will be asked some questions about the tasks and systems at the end of the experiment.

All information which is collected about you during the course of this research will be kept strictly confidential. You will be identified by an ID number and all information about you will have your name and contact details removed so that you cannot be recognised from it. Data will be stored only for analysis, then destroyed.

The results of this study will be used for my Ph.D. research. The results are likely to be published in late 2004 and will be available online at http://www.dcs.gla.ac.uk/~whiter/study/. You can request a summary of the results in the consent form. You will not be identified in any report or publication that arises from this work.

This research is being funded by the Research Student Committee at the Department of Computing Science, University of Glasgow and the Engineering and Physical Sciences Research Council (http://www.epsrc.ac.uk). This project has been reviewed by the Faulty of Information and Mathematical Sciences Ethics Committee.

For further information about this experiment please contact:

Ryen W. White (e.mail: ryen@dcs.gla.ac.uk or tel: 0141 330 2788).
Department of Computing Science, University of Glasgow
17 Lilybank Gardens
Glasgow, G12 8RZ.

Department: *Computing Science*
Subject Identification Number for this study:

UNIVERSITY
of
GLASGOW

CONSENT FORM

Title of Project:

Web Search Interface Investigation

Name of Researcher:

Ryen W. White

Please initial box

1. I confirm I have read and understand the information sheet dated (…./…./2004) (version ....) for the above study and have had the opportunity to ask questions. □

2. I understand that my permission is voluntary and that I am free to withdraw at any time, without giving any reason, without my legal rights being affected. □

3. I agree to take part in the above study. □

4. I would like to receive a summary sheet of the experimental findings □

   IF YOU WISH A SUMMARY, leave an email address ______________________

________________________	_____________	_____________________
Name of subject	Date	Signature
________________________	_____________	_____________________
Researcher	Date	Signature

1 for subject; 1 for researcher

Department: *Computing Science*
Subject Identification Number for this study:

UNIVERSITY
of
GLASGOW

RECEIPT OF PAYMENT

Title of Project:

Web Search Interface Investigation

Name of Researcher:

Ryen W. White

I confirm receipt of £12 paid for my participation in the above experiment.

Name of subject Date Signature

Researcher Date Signature

ENTRY QUESTIONNAIRE

UNIVERSITY
of
GLASGOW

This questionnaire will provide us with background information that will help us analyse the answers you give in later stages of this experiment.

ID:		System:		Task:	

Please place a TICK ☑ in the square that best matches your opinion

Section 1: PERSONAL DETAILS

1. Please provide your AGE: □

2. Please indicate your GENDER:

Male... □ 1

Female....................................... □ 2

3. Please indicate the HAND YOU USE TO CONTROL THE MOUSE:

Right.. □ 1

Left... □ 2

4. Please provide your CURRENT OCCUPATION:

5. What college/university degrees/diplomas do you have (or expect to have)?

Degree:		Subject:		Date:	
Degree:		Subject:		Date:	
Degree:		Subject:		Date:	

Section 2: SEARCH EXPERIENCE

6. Overall, for how many years have you been doing online searching?

7. Do you carry out online searches at home or work?

Yes.. □ 1

No.. □ 2

If YES, how frequently?

once or twice a year	once or twice a month	once or twice a week	once or twice a day	more often
□	□	□	□	□

8. How much experience have you had:

None — A lot

Using point-and-click interfaces e.g. Macintosh, Windows............................ □ □ □ □ □

Searching on computerised library catalogues locally (e.g. in your library) or remotely (e.g. Library of Congress)............ □ □ □ □ □

Searching on World Wide Web search services (e.g. Google, AltaVista)................ □ □ □ □ □

Searching on other retrieval systems.......... □ □ □ □ □

1 2 3 4 5

(please specify which systems)......................................

9. You find what you are searching for:

Never — Always

□ □ □ □ □

1 2 3 4 5

10. Please indicate which search engines you use (mark AS MANY as apply)

Google (http://www.google.com).. □ 1

Yahoo (http://www.yahoo.com)... □ 2

AltaVista (http://www.altavista.com).. □ 3

AlltheWeb (http://www.alltheweb.com)...................................... □ 4

Others (please specify)...... 5

11. Using the search engines you chose in question 10 is GENERALLY:

	1	2	3	4	5	
easy	□	□	□	□	□	difficult
stressful	□	□	□	□	□	relaxing
simple	□	□	□	□	□	complex
satisfying	□	□	□	□	□	frustrating

SEARCH QUESTIONNAIRE

UNIVERSITY
of
GLASGOW

To evaluate the system, we now ask you to answer some questions about it and your search in general. Take into account that we are interested in knowing your opinion: answer questions freely, and consider there are no right or wrong answers.

Please remember that we are evaluating the system you have just used and not you.

ID:		System:		Task:	

Place a TICK ☑ in the square that best matches your opinion. Please answer all questions.

Section 1: SEARCH PROCESS

1. The search we asked you to perform was:

	1	2	3	4	5	
stressful	□	□	□	□	□	relaxing
interesting	□	□	□	□	□	boring
tiring	□	□	□	□	□	restful
easy	□	□	□	□	□	difficult

Section 2: SUPPORT

Each of the three systems has different features to help you find relevant information. In this section we ask you about the system you have just used.

Content Presentation

2. As I searched, I tried to only view information related to the search task

Agree				Disagree
□	□	□	□	□
1	2	3	4	5

3. The information laid out on the results page was:

	1	2	3	4	5	
unhelpful	□	□	□	□	□	helpful
useful	□	□	□	□	□	not useful
ineffective	□	□	□	□	□	effective
not distracting	□	□	□	□	□	distracting

Choosing Additional Query Words

Each system offered terms that could be used to construct a new query for reordering sentences and documents, or re-searching the Web.

4. I felt comfortable with how the new query was constructed

Disagree				Agree
□	□	□	□	□
5	4	3	2	1

Choosing Action

Each system allowed the reordering of sentences and documents, or re-searching the Web. In this questionnaire we call this the 'action'.

5. I felt comfortable with how the action was selected

Agree				Disagree
□	□	□	□	□
5	4	3	2	1

Relevance Assessment

The Automatic and Interactive systems assumed that much of the information you viewed was relevant. In the Checkbox system you explicitly marked relevant items.

6. How you conveyed relevance to the system (i.e. ticking boxes or viewing information) was:

	1	2	3	4	5	
difficult	□	□	□	□	□	easy
effective	□	□	□	□	□	ineffective
not useful	□	□	□	□	□	useful

7. How you conveyed relevance to the system made you feel:

	1	2	3	4	5	
comfortable	□	□	□	□	□	uncomfortable
not in control	□	□	□	□	□	in control

ONLY COMPLETE 'Notification that Action has Occurred' IF YOU HAVE JUST USED THE AUTOMATIC OR INTERACTIVE SYSTEMS

Notification that Action has Occurred

8. The system communicated its action in a way that was:

	1	2	3	4	5	
unobtrusive	□	□	□	□	□	obtrusive
uninformative	□	□	□	□	□	informative
timely	□	□	□	□	□	untimely

9. The appearance of the 'idea bulb' when the system chose/recommended an action was:

	1	2	3	4	5	
not disruptive	□	□	□	□	□	disruptive
not useful	□	□	□	□	□	useful

Section 3: ADDITIONAL WORDS CHOSEN/RECOMMENDED BY THE SYSTEM

The systems chose or recommended additional query words. In this section we ask you about this process.

10. The words chosen/recommended by the system were:

	1	2	3	4	5	
irrelevant	□	□	□	□	□	relevant
useful	□	□	□	□	□	not useful

ONLY ANSWER QUESTION 11. IF YOU HAVE JUST USED THE CHECKBOX OR INTERACTIVE SYSTEMS

11. You accepted any recommended words because (mark AS MANY as apply):

they meant the same.. □ 1

they were related to words you had chosen already.............................. □ 2

you couldn't find better words... □ 3

they represented new ideas (i.e. not part of your original request)......... □ 4

other (please specify)..................... □ 5

12. The extra words ENTERED BY YOU originated in ideas from (mark ONE only):

a. the list of terms suggested by the system... □ 1

b. the retrieved set of documents and extracted information................. □ 2

c. a combination of 'a' and 'b'... □ 3

d. other (please specify)................. □ 4

13. I would trust the system to choose words for me

Agree Disagree

□	□	□	□	□
1	2	3	4	5

14. Do you have any further comments about the words that were chosen/recommended?

Section 4: ACTION CHOSEN/RECOMMENDED BY THE SYSTEM/YOU

The Automatic and Interactive systems attempt to choose actions that reflect changes in the required information. The Checkbox system lets you choose the action In this section we ask for your views on this process.

ONLY ANSWER QUESTIONS 15. to 18. IF YOU HAVE JUST USED THE AUTOMATIC OR INTERACTIVE SYSTEMS

15. The action chosen/recommended by the system reflected changes in the information you searched for:

Never — Always

□	□	□	□	□
5	4	3	2	1

16. The action chosen/recommended by the system was:

	1	2	3	4	5	
useful	□	□	□	□	□	not useful
unhelpful	□	□	□	□	□	helpful
appropriate	□	□	□	□	□	inappropriate

17. You accepted any chosen/recommended actions because (mark AS MANY as apply):

they matched what you wanted to do (i.e. were appropriate).............. □ 1

they were worth trying (i.e. to see what would happen).......................... □ 2

you hadn't considered doing them... □ 3

other (please specify).................... □ 4

18. I would trust the system to choose an action for me

Disagree — Agree

□	□	□	□	□
5	4	3	2	1

ONLY ANSWER QUESTION 19. IF YOU HAVE JUST USED THE CHECKBOX OR INTERACTIVE SYSTEMS

19. YOU CHOSE any actions because (mark AS MANY as apply):

they matched what you wanted to do (i.e. were appropriate)............... □ 1

they were worth trying (i.e. to see what would happen).......................... □ 2

similar actions had been beneficial before.. □ 3

other (please specify).................... □ 4

20. Do you have any further comments about the action chosen/recommended by the system?

□

Section 5: TASK

In this section we ask about the search task you have just attempted.

21. You chose this task because (mark ONE only) :

you had an interest in it.. □ 1

you were familiar with similar tasks.. □ 2

there were no other tasks you could do.. □ 3

it was the least boring.. □ 4

no reason.. □ 5

other (please specify).................... □ 6

22. The task we asked you to perform was:

	1	2	3	4	5	
unclear	□	□	□	□	□	clear
simple	□	□	□	□	□	complex
unfamiliar	□	□	□	□	□	familiar

23. I encounter a task similar to this one frequently

Agree Disagree

□	□	□	□	□
1	2	3	4	5

24. I had an exact idea of the type of information I wanted

Disagree Agree

□	□	□	□	□
5	4	3	2	1

25. I believe I have succeeded in my performance of this task

Agree Disagree

□	□	□	□	□
1	2	3	4	5

26. I think there was better information available (that the system did not help me find)

Disagree Agree

□	□	□	□	□
5	4	3	2	1

27. Do you have any further comments about the task you have just attempted?

EXIT QUESTIONNAIRE

The aim of this experiment was to investigate the relative effectiveness of three different Web search interfaces.

ID:		System:		Task:	

UNIVERSITY
of
GLASGOW

Please answer the following questions as fully as you feel able.

Section 1: SYSTEM EXPERIENCES

1. Rank the systems in order of preference (1 = best, 3 = worst):

Checkbox:
Recommendation:
Automatic:

2. Explain your ranking in the previous question

3. How did you feel about each system you used?
[please refer to printed screenshots if necessary]

Checkbox

Interactive

Automatic

Section 2: TASK EXPERIENCES

4. Rank the tasks in order of preference (1 = best, 3 = worst):

First Task:
Second Task:
Third Task:

5. Explain your ranking in the previous question

Section 3: COMMENTS

6. Do you have any further comments or questions about the systems or experiment?

Please take note of my email address and let
me know if you have any further questions.

Thank you for your help

Department: *Computing Science*

TASK A

Title of Project:

Web Search Interface Investigation

Name of Researcher:

Ryen W. White

UNIVERSITY of GLASGOW

Please choose one task from the six topics given below. *You may not choose a task from the same topic as any chosen previously*. You have 15 minutes to attempt this task. Please remember that it is the systems that are under evaluation, not you.

Topic

1 A friend who has been attempting to gain a university place has been complaining that there are too many people attending university today, you were unsure if this assessment was correct and have decided to find out what changes there have been in the student population in recent times.

2 You are currently working as a research assistant at a local university. Your laboratory is in an old building and one of your colleagues has developed a severe allergy that you believe is caused by his working environment. You want to gather information on allergies in the workplace that will help you advise him.

3 You are about to depart on a short-tour along the west coast of Italy. The agenda includes a visit to the country's capital, Rome, during which you hope to view many of the city's modern art galleries and museums. You decide to find out from a number of sources which are the most popular art galleries and museums, and for what reasons.

4 Whilst in a mobile phone shop, you overhear a staff member telling one of their friends to buy a 3rd Generation phone. Your friend didn't want to be sucked into buying something that may soon be obsolete so has asked you to explain 3rd Generation mobile phone technology to them.

5 Your friend has just finished reading a copy of a national newspaper in which there is mention Internet music piracy. The article stresses how this is a global problem and affects compact disc sales worldwide. Unaware of the major effects you decide to find out how and why music piracy influences the global music market.

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

Department: *Computing Science*

UNIVERSITY *of* GLASGOW

TASK B

Title of Project:

Web Search Interface Investigation

Name of Researcher:

Ryen W. White

Please choose one task from the six topics given below. *You may not choose a task from the same topic as any chosen previously*. You have 15 minutes to attempt this task. Please remember that it is the systems that are under evaluation, not you.

Topic

1 A friend has recently been applying to various universities and courses but has been complaining that they are finding it difficult to attain a place due to a much larger and varied number of people attending university. You were unaware if their assessment was correct so you have decided to find out how the composition of the student population has changed over the past 5 years.

2 You are currently working as a research assistant at a local university. A colleagues has recently been diagnosed with a dust allergy caused by dust in his working environment. He is writing a letter to the university complaining about the lack of cleanliness. He has asked for you to help him find information on the causes of dust allergies that may be useful for constructing this letter.

3 You are about to depart on a short-tour along the west coast of Italy. The agenda includes a visit to the country's capital, Rome, during which you hope to find time to pursue your interest in modern art. However, you have recently been told that time in the city is limited and you want information that allows you to choose a gallery to visit.

4 Whilst in a mobile phone shop, you overhear a staff member telling one of their friends to wait until 3rd Generation phones are available before purchasing a new one. The staff are looking for a quick sale and don't seem to be very forthcoming with information on this technology so you decide to find out for yourself what special features will be available on 3rd Generation mobile phones before making a decision.

5 Your friend has just finished reading a copy of a national newspaper in which there is mention of Internet music piracy. This article suggests that the costs of steps taken to stop the illegal downloading of music are passed directly to the consumer. You decide to research which actions have been most cost-effective in combating the problem.

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

Department: *Computing Science*

UNIVERSITY
of
GLASGOW

TASK C

Title of Project:

Web Search Interface Investigation

Name of Researcher:

Ryen W. White

Please choose one task from the six topics given below. *You may not choose a task from the same topic as any chosen previously*. You have 15 minutes to attempt this task. Please remember that it is the systems that are under evaluation, not you.

Topic

1 A friend has recently been applying to various universities and courses but has been complaining that they are finding it difficult to attain a place due to the rising numbers of students. You were unsure if their assessment was correct so you have decided to find out how the size of the student population changed over the last 5 years and how it is expected to change in the coming 5 years.

2 You are currently working as a research assistant at a local university. Your laboratory is in an old building and one of your colleagues has recently been diagnosed with a dust allergy caused by dust in his working environment. You decide to help him by finding some simple steps that can be taken to tackle dust allergies.

3 You are about to depart on a short-tour along the west coast of Italy. The agenda includes a visit to the country's capital, Rome, during which you want to want to visit an art gallery. Your friend has an interest in impressionist paintings and you would like to find a gallery in Rome that has such paintings.

4 Whilst in a mobile phone shop, you overhear a staff member telling one of their friends to wait until 3G or 3rd Generation phones are available before purchasing a new one. The staff are looking for a quick sale and don't seem to be very forthcoming with information on this technology so you decide to find out for yourself what special features will be available on 3G or 3rd Generation mobile phones before making a decision.

5 You are having a discussion with your friend about an article on Internet music piracy. Your friend suggests that illegal music downloads are affecting sales of compact discs, and driving up compact disc prices in Europe in particular. Unsure if this is true, you decide to find out whether music piracy has a direct influence on European compact disc prices, and if so, to what extent.

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

Department: *Computing Science*

UNIVERSITY
of
GLASGOW

TRAINING TASK

Title of Project:

Web Search Interface Investigation

Name of Researcher:

Ryen W. White

Please read the task description below and once you feel comfortable that you understand what is required, try using the training system to attempt it.

> Next weekend, a close friend of yours is hoping to go on a short-break to Paris, France. He has recently moved house and does not have a phone line installed. As a result he has asked you to look for hotels in the city on his behalf. Both of you are not too confident with your French speaking skills and would like to find hotels that offer an online registration service. Your friend expects to get Internet access again soon and he would like the Web address (e.g., http://...) from five such hotels in the city, so that he can pursue the booking himself.

Department: *Computing Science*

UNIVERSITY
of
GLASGOW

TASK ANSWERS/NOTES

Title of Project:

Web Search Interface Investigation

Name of Researcher:

Ryen W. White

ID:		System:		Task:	

Please write your answers or any notes in the space provided below. If you require more paper, please ask the experimenter. You have 15 minutes to attempt this task.

Appendix G

In this Appendix I present details of the Interaction logs created during the experiment. The tags used in the log files are given in Appendix G.1 and an excerpt from the logs is included in Appendix G.2.

Appendix G.1

The tags used in the interaction logs are described in the tables below. The symbol '#' is used to represent a number where appropriate.

Table G.1.
General interaction tags.

Tag	Meaning
SENT[doc #][sentence #]	Sentence clicked
LRSENT[sentence #]	Low-ranked sentence clicked (rank above 15)
SENTAR[sentence #]	Sentence arrow clicked
L[#]	Length of top-ranking sentence list
DOC[doc #]	Document viewed
LRDOC[doc #]	Low-ranked document viewed (beyond first 10)
HIGHDOC[doc #]	Document title highlighted
LRHIGHDOC[doc #]	Low-ranked document highlighted
SUM[doc #]	Summary viewed
SUMFAIL[doc #]	Summary could not be created because of technical problems
SUMOK[doc #]	Summary created
SS[doc #][sentence #]	Summary sentence clicked
SIC[doc #][sentence #]	Sentence-in-context viewed
NEXT[start #]	Next button clicked
PREV[start #]	Previous button clicked
NP	New relevance path
STEP[#]	Step number in relevance path
COORD[#,#]	Position of the mouse pointer [x-coordinate, y-coordinate]
CLICKCOORD[#,#]	Position of a mouse click [x-coordinate, y-coordinate]

Table G.2.
Explicit relevance assessments tags.

Tag	Meaning
XTITLE[doc #]	Title relevant
XSENT[sentence #]	Top-Ranking Sentence relevant
XSUM[doc #]	Summary relevant
XSS[doc #][sentence #]	Summary sentence relevant
XSIC[doc #][sentence #]	Sentence in context relevant
XCA	Clear all checked representations

Table G.3.
Result set information tags.

Tag	Meaning
RESREP[#]	Total number of potential representations
RESDOC[#]	Total number of documents returned

Table G.4.
Queries and query modification tags.

Tag	Meaning
Q[$t_1,\ldots,t_n$]	Original query
EC[rank position #][*t*]	Expansion term chosen from list of potential expansion terms
ER[rank position #][*t*]	Term removed from list of chosen expansion terms
EL[#]	Expanded query length
EXP[$t_1,\ldots,t_n$]	Expanded query
ECA	Clear all expansion terms
XCQ	Clear query from Checkbox system
XRQ	Restore query from Checkbox system
XTA[*t*]	Add term *t* from Checkbox system
XTD[*t*]	Remove term *t* from Checkbox system

Table G.5.
Retrieval strategy (action) tags.

Tag	Meaning
AU	Undo action
AU-MIN	Undo action from minimised window (Automatic system)
AU-MAX	Undo action from maximised window (Automatic system)
AREC[a]	Recommended action

Appendix G.2

An excerpt from the interaction logs of the Recommendation system for the search task on dust allergies. The initial query was ‘causes dust allergy’ and the contents of the EXP[..] tag are the top 20 terms recommended by the system.

```
COORD[585,401] : 1078940148799 : Wed Mar 10 17:35:48 GMT 2004
COORD[569,400] : 1078940149050 : Wed Mar 10 17:35:49 GMT 2004
COORD[601,396] : 1078940149310 : Wed Mar 10 17:35:49 GMT 2004
COORD[620,395] : 1078940149560 : Wed Mar 10 17:35:49 GMT 2004
COORD[557,404] : 1078940149821 : Wed Mar 10 17:35:49 GMT 2004
COORD[589,408] : 1078940150071 : Wed Mar 10 17:35:50 GMT 2004
COORD[572,404] : 1078940150321 : Wed Mar 10 17:35:50 GMT 2004
COORD[571,403] : 1078940150572 : Wed Mar 10 17:35:50 GMT 2004
SENTAR[8] : 1078940150692 : Wed Mar 10 17:35:50 GMT 2004
EXP[house allergic information medical mite faq treatment medication
options learn reasons advice allergies symptoms asthma health mold
allergens pollen air] : 1078940150882 : Wed Mar 10 17:35:50 GMT 2004
COORD[571,403] : 1078940150942 : Wed Mar 10 17:35:50 GMT 2004
AREC[trs] : 1078940150952 : Wed Mar 10 17:35:50 GMT 2004
NP[21] : 1078940150962 : Wed Mar 10 17:35:50 GMT 2004
STEP[21][1] : 1078940150962 : Wed Mar 10 17:35:50 GMT 2004
LRHIGHDOC[29] : 1078940150962 : Wed Mar 10 17:35:50 GMT 2004
CLICKCOORD[571,403] : 1078940151002 : Wed Mar 10 17:35:51 GMT 2004
COORD[571,403] : 1078940151253 : Wed Mar 10 17:35:51 GMT 2004
COORD[440,404] : 1078940151513 : Wed Mar 10 17:35:51 GMT 2004
COORD[375,400] : 1078940151763 : Wed Mar 10 17:35:51 GMT 2004
COORD[350,390] : 1078940152014 : Wed Mar 10 17:35:52 GMT 2004
COORD[350,392] : 1078940152264 : Wed Mar 10 17:35:52 GMT 2004
COORD[350,394] : 1078940152525 : Wed Mar 10 17:35:52 GMT 2004
TDOC[29] : 1078940152595 : Wed Mar 10 17:35:52 GMT 2004
DOC[29] : 1078940152595 : Wed Mar 10 17:35:52 GMT 2004
TDOC[29] : 1078940152595 : Wed Mar 10 17:35:52 GMT 2004
STEP[29][2] : 1078940152785 : Wed Mar 10 17:35:52 GMT 2004
CLICKCOORD[350,394] : 1078940152785 : Wed Mar 10 17:35:52 GMT 2004
COORD[350,394] : 1078940152795 : Wed Mar 10 17:35:52 GMT 2004
COORD[738,234] : 1078940153055 : Wed Mar 10 17:35:53 GMT 2004
COORD[711,13] : 1078940153306 : Wed Mar 10 17:35:53 GMT 2004
COORD[955,164] : 1078940153766 : Wed Mar 10 17:35:53 GMT 2004
COORD[932,137] : 1078940154017 : Wed Mar 10 17:35:54 GMT 2004
COORD[787,161] : 1078940154267 : Wed Mar 10 17:35:54 GMT 2004
COORD[640,161] : 1078940154527 : Wed Mar 10 17:35:54 GMT 2004
COORD[590,127] : 1078940154778 : Wed Mar 10 17:35:54 GMT 2004
COORD[521,108] : 1078940155038 : Wed Mar 10 17:35:55 GMT 2004
COORD[514,101] : 1078940155288 : Wed Mar 10 17:35:55 GMT 2004
COORD[514,101] : 1078940155549 : Wed Mar 10 17:35:55 GMT 2004
COORD[514,101] : 1078940155799 : Wed Mar 10 17:35:55 GMT 2004
```