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.