Intertemporal Choice for Browsing in Information Retrieval

Azreen Azman

Department of Computing Science
Faculty of Information and Mathematical Sciences
University of Glasgow

UNIVERSITY
of
GLASGOW

Submitted for Degree of Doctor of Philosophy

at the University of Glasgow

January 2007

To my parents and my wife

Acknowledgements

It took me four years to complete my Ph.D and it has been one the most fulfilling experience for me. The journey will forever be remembered in my life.

I would like to thank my supervisors, Dr. Iadh Ounis and Prof. Keith van Rijsbergen for their guidance, for their inspiration and motivation and for their support throughout my four years stay in the department. Without them, I may not survive and I may not have the chance to write this thesis. I would also like to thank the members of the Information Retrieval Group for your friendship and your support. My special thank is to Jana and Sachi for their support that motivates me to complete my Ph.D. Your friendship will be forever remembered.

I want to express my gratitude to my funding body, Majlis Amanah Rakyat (MARA), and the government of Malaysia for giving me this opportunity to pursue my dream and also for the support throughout the four years of my Ph.D study.

To all my friends in Glasgow, thank for your friendship and your time. You have made my short stay in Glasgow enjoyable. I really appreciate your supports and your understanding.

To my mother, thank you very much for your understanding, your patience and you love for me. It will be impossible for me to reach to this stage of my life without your support. This Ph.D is for you, mother. To all my brothers and sisters, thank you for believing in me.

To my wife, Zazuneezan, thanks for your patience and your motivation that make it easier for me to go through this process. I love you.

Declaration

I hereby declare that this thesis was composed by me based on my own work except where explicitly stated otherwise in the text, and to the best of my knowledge it contains no plagiarised material. This work has not been submitted for any degree or professional qualification except as specified. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

Azreen Azman

Abstract

Browsing is an important strategy for information seeking in information retrieval (IR). Usually, browsing is guided by the information need of a user, where the documents are chosen by anticipating whether they could satisfy the user's information need. Therefore, the effectiveness of browsing depends on the ability of the user to make the right decision. However, user is unfamiliar with the document collection and the models underlying the IR system. Due to this limitation, the user is unlikely to make an optimal decision for his/her browsing strategy.

Such a problem can be overcome by incorporating a recommendation model to suggest a good browsing strategy for the user. A good recommendation model should be based on modelling the decision behaviour of the user. However, modelling such behaviour is problematic. In this thesis, the *intertemporal choice model* is adopted to model the browsing behaviour of the user. It is based on the assumption that browsing is an intertemporal choice problem.

The effectiveness of modelling the browsing behaviour of the users is evaluated in the context of browsing on mobile devices and post retrieval browsing. First, an implicit RF system is proposed for mobile devices to overcome the limitations of the devices, namely the small screen size and the limited interaction capability. A number of implicit RF models and display strategies are investigated to find the optimal setting for the system. The results suggest that the implicit RF system can be effective provided that an effective browsing recommendation model is incorporated. For this purpose, a recommendation system based on the intertemporal choice model is proposed. The effectiveness of the model is measured by the median average precision (MAP) and the expected search length (ESL) to measure the cost of browsing of the recommended browsing strategy.

Second, the effectiveness of the model is evaluated for post retrieval browsing in the context of the *subtopic relevance retrieval* application. Post retrieval browsing refers to the sequential assessment of the top retrieved documents. In this context, a topic consists of a set of subtopics and a document can be relevant to one and up to all subtopics. The aim of the model is to produce a ranking such that it will take as little as possible to cover all relevant subtopics by browsing through the ranks of the documents.

The results from both evaluations suggest that the intertemporal choice model could be effective provided that the parameters associated with the model are optimised and the value of the documents used in the model is as accurate as possible.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

1.1 Introduction

Browsing is regarded as an important strategy for information seeking (Bates, 1989), especially in an interactive information retrieval (IR) setting, such as on the Web, or in an implicit relevance feedback (RF) system (Vinay et al., 2005, White et al., 2005). While browsing, the user moves from one document to another to find the information he/she is looking for. Usually, browsing is guided by the information need of the user (Armstrong et al., 1995, Chi et al., 2001) such that the next document to be browsed is chosen by anticipating that the document will satisfy his/her information need. Browsing eliminates the need for the user to formulate a query representing his/her information need, which is the case in the query-based searching strategy.

However, browsing is considered to be an inefficient information seeking strategy (Olston and Chi, 2003) because it relies on the user to make a decision on the selection of documents to be browsed in a browsing session. Usually, the documents selected by the user in the previous stage of browsing will determine the documents to be offered in the next iteration. For an implicit RF system, the selected documents are used as evidence to formulate a *refined query* to retrieve another set of documents in the next iteration. In the case of the Web, the candidate documents to be selected in the next iteration are those linked to the currently visited document. Therefore, a mistake in selecting a wrong document may result in an unsuccessful browsing session. Although backtracking is possible on the Web, it is troublesome and inefficient.

There have been many attempts to assist the user while browsing on the Web. For instance, the *anchor text* of the hyperlink is used as an indication of the content of the linked document (Chi et al., 2001) to help the user in choosing the right document to proceed. In addition, links can be recommended to the user by learning from his/her browsing history (Armstrong et al., 1995, Lieberman, 1995, Mobasher et al., 2002, Olston and Chi, 2003).

However, the stage-by-stage recommendation of documents or links may not guarantee an optimal browsing session for the user. Such a recommendation strategy is unable to anticipate all possible selection strategies of the user. Moreover, it cannot look beyond the current stage of browsing to anticipate the possible results of the browsing session. An example of the problem will be discussed in Chapter 4 in the case of the implicit relevance feedback (RF) system. Due to this limitation, this thesis proposes a browsing recommendation model to encourage an effective browsing session for the user.

The *information foraging theory* suggests that the user modifies his/her information seeking strategy or modifies the structure of the information environment to maximise his/her *rate* of getting the relevant information (Pirolli and Card, 1999). Based on the theory, it is understood that an effective browsing path should be determined by modelling the browsing behaviour of the user, such as modelling the decision making behaviour of the user while browsing.

However, modelling decision behaviour of users is difficult. It requires a lot of data concerning how the user makes decisions while browsing. Web access logs have been used to model the browsing behaviour of users (Chen et al., 1998, Perkowitz and Etzioni, 2000, Mobasher et al., 2001, Mobasher et al., 2002). The logs consist of the information about the documents visited by the users in a website. The browsing paths used by the users can easily be extracted from the logs. However, due to the lack of real search tasks and relevance judgments, it is difficult to make sense of the data. In addition, there is no indication of a success or a failure for those browsing paths. Therefore, the information from Web access logs may not be useful for modelling the browsing behaviour of the user.

The *intertemporal choice* model (Read, 2004) is used to capture human decision behaviour on the choices for the rewards received or expenses paid at different times (or in succession), such as to invest a smaller amount of money now for a larger amount of money received after a year. The model is extensively investigated for the intertemporal choice problem in economic and social studies (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Frederick et al., 2002, Lazaro et al., 2002).

In this thesis, the intertemporal choice model is adopted to model the user's browsing behaviour. It is based on the assumption that browsing is an example of an intertemporal choice problem. The problem of choosing a browsing path is treated as the problem of finding the best sequence of documents for the user. In browsing, it is assumed that a user receives the documents (or rewards) each at a different time. Furthermore, it is assumed that the behaviour of the user in choosing the sequence of documents follows the intertemporal choice model. Therefore, this thesis suggests that the best sequence of documents for the user can be determined by using the intertemporal choice model.

The effectiveness of the intertemporal choice model is investigated for browsing in two different scenarios, browsing in an interactive implicit RF system and in *post-retrieval* browsing. First, the problem of browsing in an interactive implicit RF system for mobile devices is investigated (Vinay et al., 2005). The system displays *four* documents each time and it allows the user to browse by selecting one of the documents displayed to him/her. The documents selected in the previous iterations determine the documents to be retrieved in the next iteration. The aim of the model is to identify the sequence of documents that should be chosen by the user such that more relevant documents can be found with the least effort from him/her.

Second, the problem of post-retrieval browsing is investigated, which is to find an optimal strategy for assessing the top-n retrieved documents. In particular, the top-n retrieved documents are re-ranked such that the effectiveness of browsing the documents by following the ranks is optimised. It is investigated in the context of the *subtopic relevance retrieval* application (Zhai et al., 2003). In this application, a set of subtopics is defined for each topic and each document is assessed based on its relevance to each subtopic. A document can be relevant to any of the subtopics and a document is assumed to be more relevant if it contains more relevant subtopics. The documents

should be ordered such that the next document to be read should contain more *novel* information.

1.2 The implicit relevance feedback system for mobile devices

The purpose of an IR system is to retrieve the documents that are relevant to the information need of the user. The query submitted to the system is assumed to be the best representation of the user's information need. It is the responsibility of the user to choose the right keywords in his/her query. The system interprets the information need of the user based on the query and makes predictions on the documents in its collection that could satisfy the information need. Those documents are returned to the user for assessment. The user will find that his/her information seeking session is successful if the returned documents contain the information he/she is looking for.

Usually, the search session of the user involves multiple iterations of retrieval by re-formulation and re-submission of the query or by an interactive browsing with the system. The information need of a user is usually vague (Spink et al., 1998). Moreover, the user is unfamiliar with the document collection underlying the IR system (Furnas et al., 1987, Salton and Buckley, 1990). Therefore, the *query formulation* process is a problematic situation for the user. Furthermore, the user often uses a short query that consists of only a few words or query terms (Silverstein et al., 1999). Those are probably some of the reasons why the query chosen by the user may fail to retrieve relevant documents.

An interactive IR (IIR) system could be a better solution for the query formulation problem. An interactive query expansion application that is based on relevance feedback (RF) could improve the query of the user by learning from his/her feedback (Ruthven, 2003, White et al., 2005). The RF technique has been shown to be effective in a non-interactive environment (Salton and Buckley, 1990) and in a limited interactive environment (Beaulieu, 1998). The usefulness of an IIR system with RF has already been recognised (Koenemann and Belkin, 1996).

In the context of RF, it is assumed that the user has an information need and the information need is consistent throughout the iterative RF process. Based on the user's relevance judgement on the documents returned by the system, the initial query of the

user is modified such that it will include more relevant terms or the weight of the terms in the query is adjusted accordingly. The representation of the information need is assumed to be improved after multiple RF iteration as the query is continuously modified based on the feedback of the user. Based on this assumption, the effectiveness of the IR system will improve by the use of RF (Salton and Buckley, 1990).

One of the main challenges of an interactive RF system is to encourage the user to provide explicit feedback to the system. The need to assess a list of documents and eventually to provide an indication of which documents are relevant from the list is a burden to the user. An implicit RF system, on the other hand, treats the interaction of the user as implicit RF evidence. Some of the measures, such as *reading duration* and *action of saving or printing a document*, are used to imply the usefulness of the documents (Kelly and Teevan, 2003). Accordingly, the relevance of the documents for RF is implied based on those measures.

In the case of an implicit RF system for mobile devices (Vinay et al., 2005), only a small number of documents is returned to the user for investigation. Of those documents, the user selects one document that could contain the information he/she is looking for. Such an action is treated as an implicit indication of relevance for that document. Accordingly, the evidence is used by the RF model to infer the current information need of the user, which is eventually used to retrieve a better set of documents in the next iteration. As a result, the user is browsing through the collection, where at each step he/she needs to decide which document to choose. Therefore, the effectiveness of the browsing session depends on the documents selected by the user during the browsing session. It is the aim of the intertemporal choice browsing model to predict the most effective browsing strategy to be used in the implicit RF system.

1.3 The subtopic relevance retrieval application

Subtopic relevance retrieval is a special application where a set of subtopics is defined for a given topic (Zhai et al., 2003). A document can be relevant to none and up to all of the subtopics, as opposed to the conventional topical relevance retrieval where a document is assessed for relevance to a topic (Zhai et al., 2003). The main challenge for this application is to produce an effective ranking of documents such that it will promote *novelty* if the documents are browsed by following the ranks. Novelty refers to

the scenario where the next document to be browsed contains the subtopics that have not been covered by the previous documents.

The *Maximal Marginal Relevance* (MMR) ranking method (Carbonell and Goldstein, 1998) was proposed as a solution to produce an effective ranking for the subtopic relevance retrieval application (Zhai et al., 2003). The method assumes that the usefulness of a document to the user depends on its relevance to his/her query and its novelty with respect to the previously read documents. Each retrieved document is assigned with an MMR score, in which the score is a combination of the relevance value and the novelty value of the document. First, the most relevant document is assigned to the top rank. Then, the MMR score for the remaining documents are computed, where the novelty score for each document is estimated with respect to the document(s) already in the ranking. The document with the highest MMR score is assigned to the next rank. The process continues until all the documents are ranked.

As a result, the best strategy to browse the top-ranked documents for the subtopic relevance retrieval application is to follow the ranks. However, the MMR ranking method considers only one way to rank the documents, which is by iteratively selecting the documents with the highest MMR score. Since the novelty of a given document depends on the documents read previously, different ordering of documents may produce a different novelty value for the document. Therefore, a different MMR score could be assigned to the document. In such cases, there is a chance that the ranking produced by the MMR ranking method is not the optimal ranking for the subtopic relevance retrieval application.

In this thesis, the problem of finding the optimal ranking of the top-n retrieved documents for the subtopic relevance retrieval application is investigated. The aim is to choose the best ordering of the documents from all possible ordering of those documents. Again, the intertemporal choice model is used to find the best sequence of documents to support effective browsing.

1.4 Thesis statement

This thesis argues that the *incorporation of an intertemporal choice model into an IR system can improve the effectiveness of browsing for the user using the system.* It is investigated in the context of browsing for an implicit RF system and in the context of

post-retrieval browsing for the subtopic relevance retrieval application. The main strength of the intertemporal choice model is its capability to model the decision behaviour of an individual in the context of the intertemporal choice problem. In particular, it can measure the usefulness of a sequence of outcomes to an individual. Assuming that browsing is an intertemporal choice problem and the outcomes are the documents; the model will be able to make predictions on the best browsing strategy for the user of an IR system.

1.5 Thesis outline

The remainder of this thesis is organized in the following way.

Chapter 2: *Background*

Chapter 2 provides an overview of the research related to the work carried out in this thesis. The discussion focuses mainly on the issues in browsing and interactive IR. Moreover, the thesis provides a detailed discussion on various issues of the intertemporal choice model in the context of economic and social studies.

Chapter 3: *The Intertemporal Choice Browsing Model*

In Chapter 3, an intertemporal choice browsing model for IR is proposed. The model attempts to provide a better prediction technique to choose the best browsing strategy for the user that involves a sequential assessment of documents. It is based on the assumption that the browsing problem is an intertemporal choice problem (Read, 2004). The proposed model is motivated by the success of the intertemporal choice model in the area of economic and social studies (Loewenstein and Prelec, 1992, Lazaro et al., 2002). The effectiveness of the proposed model is evaluated as a browsing recommendation model for an implicit RF system (Chapter 5) and as an optimal ranking model for the subtopic relevance retrieval application (Chapter 6).

Chapter 4: *An Implicit Relevance Feedback System for Mobile Devices*

Chapter 4 investigates the effectiveness of an implicit relevance feedback system for mobile devices. The aim of the system is to overcome two main limitations of mobile devices, their small screen and their limited interaction capability. Such a retrieval strategy was first proposed in (Vinay et al., 2005), which limits the number of the retrieved documents to be displayed to *four* and relies on the interaction of the user to improve the quality of the retrieved documents for the next iteration. In this chapter, the

effectiveness of the ostensive model as the implicit relevance feedback models for the system are studied. The evaluation is conducted on TREC test collections (Harman, 1993).

Chapter 5: *A Browsing Recommendation Model for the Implicit Relevance Feedback System*

As a result from the investigation in Chapter 4, the effectiveness of an implicit relevance feedback system could be improved by incorporating a browsing recommendation model to assist the user in making the decision on which documents to choose. In Chapter 5, the effectiveness of a browsing recommendation model based on the intertemporal choice browsing model proposed in Chapter 3 is investigated. In the evaluation, various assumptions of the model as stated in Chapter 3 are investigated. The evaluation is conducted on the same test collections as in Chapter 4.

Chapter 6: *Re-ranking of the Top-n Documents for Post Retrieval Browsing*

In Chapter 6, the effectiveness of the intertemporal choice browsing model is investigated for the problem of post-retrieval browsing on the top retrieved documents in the context of the subtopic relevance retrieval application (Zhai et al., 2003). In particular, the top retrieved documents are re-ranked such that a sequential assessment of the documents by following the ranks is the most effective.

Chapter 7: *Conclusion and Future Work*

Chapter 7 summarises the findings of the evaluations conducted in this thesis. Conclusions on the problems investigated in this thesis will be given. Some of the research that could be done to further investigate the outcomes of this thesis is also provided.

2 Background and Motivation

2.1 Introduction

This chapter discusses the research related to the problems investigated in this thesis. First, the general architecture of an information retrieval (IR) system is presented in Section 2.2. Two processes involved in the IR system, the indexing and the retrieval processes, are discussed. Second, the shortcomings of the general IR model are presented and the motivations for using the interactive relevance feedback (RF) approach are described in Section 2.3. In particular, this section reviews some of the techniques for implicit RF. Third, the problems and potential solutions for effective information seeking on mobile devices are discussed in Section 2.4. Fourth, Section 2.5 gives an overview of the browsing problem investigated in this thesis and some of the solutions for the problem attempted by other researchers. Fifth, the intertemporal choice model is discussed in Section 2.6 as a potential solution for the browsing problem discussed in this thesis. Finally, the summary for this chapter is given in Section 2.7.

2.2 The Architecture of an Information Retrieval System

The general model of an IR system is depicted in Figure 2-1. In a typical scenario, a user submits a request (or a query) to an IR system in order to find documents that fulfil his/her information need. The system, then, finds and retrieves the documents from the collection that it judges relevant to the user's request. The retrieved documents are

returned to the user for assessment. If the user finds the documents he/she intended from the retrieved set, the information seeking process is considered a success.

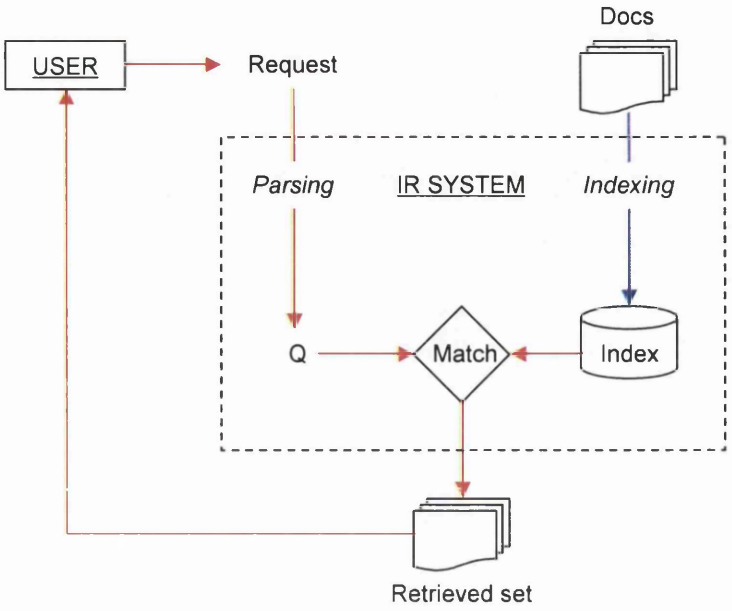


Figure 2-1: A general architecture of an information retrieval system

There are two processes involved, the indexing process indicated by the blue arrows and the retrieval process indicated by the red arrows. The indexing process takes place before the actual retrieval process. The indexing process involves transforming the documents into a representation understood by the system and suitable for retrieval. The documents' representations are stored within the IR system's repositories. The retrieval process occurs when a user submits a query to the IR system. The system parses the query and transforms it into a representation understood by the system analogous to the indexing process of documents. The query representation is matched with the documents' representations stored in the index to find the *relevant* documents. Finally, the relevant documents, as determined by the system, are retrieved and returned to the user.

2.2.1 Indexing

Indexing is a process whereby the documents in the collection are processed in order to create a suitable representation of the documents in the system. Usually, each document is represented by the *keywords* or the *terms* it comprises. However, not all the terms are significant to represent the topic or the subject of the document. One way to measure the significance of a term in a document is based on its frequency of occurrence within the document (Luhn, 1958). There are terms that occur frequently in many documents

but they are not related to the topic or the subject of the documents. These terms are called *stop words*, which are removed from the documents' representation in the index. Examples of the stop words are 'the', 'a', 'and' and so on. Another step involved in the indexing process is stemming. The aim of stemming is to reduce the term to its root form. Stemming is usually language-dependent and Porter's stemming algorithm is probably the most popular algorithm for the English language (Porter, 1980). Usually, the terms are stored as an *inverted file* for efficient retrieval. The terms stored in the inverted file are called the *index terms*.

2.2.2 Query

A user, with a particular information need, approaches an IR system to look for the documents that could satisfy his/her need for information. The information need is motivated by a gap in his/her current state of knowledge (Belkin et al., 1982), a gap between what he/she knows and what he/she wants to know. The user looks for the relevant information to close the gap. In order to satisfy the need, a user formulates a *request* in the form of a set of keywords that best represent his/her information need. This process is known as *query formulation*. The request is a specification of the information that a user is trying to find (Oddy, 1977).

The success of an IR system may depend on the quality of the request/query submitted by the user to the system. It is a quality-in, quality-out principle, where a request/query that more accurately represents the real needs of the user will have a higher chance to produce a better result (Croft and Thompson, 1987). However, the information need of the user is often vague, (Spink et al., 1998) which makes the query formulation process more problematic. Choosing the right keywords to represent the information need of the user is usually not an easy task. It becomes more problematic when the user is unfamiliar with the document collection underlying the retrieval environment (Furnas et al., 1987, Salton and Buckley, 1990). In particular, the user is not familiar with the documents in the collection, exactly how the matching of the query and the documents is achieved and which keywords are suitable to find the intended documents.

Furthermore, the user's query is often very short, containing only a small number of query terms (Silverstein et al., 1999) while there is a huge number of documents in the collection, especially on the Web. The terms/keywords in the query may fail to discriminate between relevant and non-relevant documents. On the Web, the hyperlink

structure could be used as an additional source of evidence to discriminate good documents (Brin and Page, 1998, Kleinberg, 1999). On the other hand, the query may be expanded to include more appropriate terms in order to improve the performance of an IR system. This process is known as *query expansion.* Query expansion is usually initiated by the IR system interactively, such as by suggesting a set of good query terms for the user to choose from, or automatically, such as by adding new query terms without the user consent (Ruthven, 2003). For instance, new query terms can be generated based on the *query concepts* derived from the documents in the collection (Chang et al., 2006) or through relevance feedback (Salton and Buckley, 1990, Harman, 1992, White et al., 2005). Since new query terms are generated by the IR system based on its underlying retrieval model and also based on the documents indicated to be relevant, the performance of an IR system can be improved.

2.2.3 Matching

The matching is a process of finding the documents in the collection that are relevant to the request/query of the user. Before the matching, the request of the user is *parsed,* where the stop words are removed and the remaining query terms are stemmed, following the same procedure applied during the indexing. As a result, the processed query is in the form that is understood by the system and suitable for matching.

The oldest matching function is probably the Boolean model (Van Rijsbergen, 1986b). The query for the model is expressed in Boolean form. For instance, a query '*jaguar* AND *car* AND (NOT *animal*)' is used to retrieve the documents containing the word '*jaguar*' and '*car*' and without the word '*animal*' in them. The apparent problem with the model is that the retrieved documents are not ranked and it will be difficult for the user to assess the relevance of the documents. This is one of the reasons why this model is not widely used. Another classical retrieval model is the Vector Space Model (Salton et al., 1997). It is based on the idea that both the query and the documents can be represented as vectors in the same space. The dimension of each vector corresponds to the number of the index terms in the system, or the size of the vocabulary for the document collection. The rank for each document is determined based on its similarity to the query, which is actually the distance measured between the document vector and the query vector in the hyperspace. There are several distance functions that can be used to measure similarity between the vectors (Van Rijsbergen, 1979).

Another type of matching function is the probabilistic model. The *probabilistic indexing model* proposed in (Maron and Kuhns, 1960) was probably the first probabilistic model for IR. The weight of an index term in a given document is estimated based on the probability that a user finds a given document relevant, will use the index term in his/her query. Therefore, the index term is assigned with the probabilistic weight. Meanwhile, the *probabilistic relevance model* measures the probability of relevance of a document to a given query (Robertson and Sparck Jones, 1976). The model assumes that there exists the knowledge about the distribution of the index terms in the relevant documents and through an iterative process the distribution of the index terms are refined, which makes the estimation of the probability better. Another type of the probabilistic models is the *statistical language model* for IR (Ponte and Croft, 1998). A probability value is assigned to a document based on the probability that a given query is generated from that document. Accordingly, the documents are ranked based on their probability values.

2.2.4 Evaluation

Finally, a set of potentially relevant documents to the query of the user as judged by the IR system is discovered by the corresponding matching function. The documents are returned to the user for assessment. A common strategy is to list the documents in decreasing order of their probability of relevance or their retrieval status value (RSV) for the user. The underlying idea is based on the Probability Ranking Principle (PRP) (Robertson, 1977). The formal statement of the PRP is as follows:

> *"The Probability Ranking Principle: If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data"*

According to the principle, the documents which are more relevant to the user's information need could be found at the top ranks. In the case of the probabilistic model for IR, the PRP has been argued to maximise the likelihood of finding the relevant documents (Gordon and Lenk, 1991) provided that the following three conditions are satisfied; the retrieval system is well-calibrated such that the probability estimation is as

accurate as possible to the usefulness of the document to the user, the predicted probabilities are reported with certainty and the documents are assessed for relevance independently. It has been shown in (Gordon and Lenk, 1992) that any violation of these conditions will make the ranking become sub-optimal.

In order to measure the quality of an IR system in response to a given query, the retrieved documents are assessed for relevancy. The concept of relevance is dynamic and multi-dimensional (Borlund, 2003). A document could be relevant to a given user at a given time and could be irrelevant at any other time, which reflects the dynamic nature of the concept of relevance. Moreover, a document could be perceived to be relevant to a given query by a particular user but could be irrelevant to any other user even with the same query formulated, which demonstrates its multi-dimensionality.

The evaluation design for an IR system is based on the early works in the Cranfield test (Cleverdon, 1997). The Textual Retrieval Conference (TREC) provides a more standardised evaluation platform for IR systems (Harman, 1993). TREC provides a number of test collections with a set of queries, document collections and relevance judgments for the respective queries, which enable a more standardised comparison among different IR systems. The effectiveness of an IR system is usually measured by its *precision* and its *recall* with respect to a given query (Van Rijsbergen, 1979).

2.3 Interactive Relevance Feedback System

In a typical information seeking scenario, the user is actually interacting with the IR system. The submission of a request to the system and the subsequent assessment of the retrieved documents are an example of interaction between the user and the IR system. In some cases, the user may reformulate and resubmit his/her request in order to retrieve another set of documents. Moreover, the user may start browsing by selecting one of the retrieved documents in the case of the Web. Therefore, the information seeking session is often iterative and interactive.

However, the general model presented in Section 2.2 does not capture this scenario. It has been argued in (Bates, 1989) that such a model is unable to capture the real information seeking behaviour of the user. It is important to include the user in modelling an IR system. The original query submitted to the system may not be good enough to correctly predict the relevancy of the documents. The user may not find

relevant documents in the initial retrieval iteration (Van Rijsbergen, 1986a). However, the user has the capability to identify those documents that are relevant to his/her information need. Usually, the user will interact with the IR system, by modifying his/her query in order to retrieve more relevant documents. In the case of the RF application, the system seeks feedback from the user regarding the relevancy of the retrieved documents. That information is used to modify the user's initial query in order to retrieve more documents in the next iteration. Therefore, the interaction between the user and the IR system could improve the effectiveness of the information seeking session.

Moreover, the interaction can be used to resolve ambiguity in the semantic context of the query, by, for example, using a *query-answering game* (Agostini and Avesani, 2004). It is based on the *guessing game* proposed as a part of the *language game* in (Steels, 2001) where an iterative process is used to unify the word-meaning association between two or more vocabulary sets (Steels, 2001). Such a technique is useful to resolve the ambiguity of a query term that has more than one meaning. For instance, a query '*jaguar*' may refer to an animal or a car. The system may interpret the query term as a car while the user may be looking for the documents about an animal. Therefore, the top retrieved documents may not consist of the relevant documents. By using this technique, both the user and the system should have a common vocabulary set and the ambiguity problem should be resolved. The use of this technique has been investigated in the context of *advertising games* for classifying Web directories (Agostini and Avesani, 2003) and in the peer-to-peer environment for unifying the *commands* between peers (Avesani and Agostini, 2003).

The RF technique recognises the role of the user in improving the effectiveness of the system beyond the initial query formulation. It has been suggested that the RF technique can improve the effectiveness of an IR system in a non-interactive environment (Salton and Buckley, 1990) or in a restricted interactive environment (Beaulieu, 1998). The RF technique can be applied without explicit user feedback, such as the *pseudo*-RF in which a certain number of the top retrieved documents are assumed to be relevant (Tao and Zhai, 2006). However, such an assumption is not necessarily founded. In addition, the need to specify a suitable number of the top retrieved documents is problematic. Therefore, there is a need to include the user in the process, which indicates that the RF technique should be applied in an interactive environment.

The usefulness of using the RF technique for an interactive IR system has already been recognised (Koenemann and Belkin, 1996). Research on the application of the RF technique in an interactive IR system focuses on two areas, the effectiveness of the system interface design and the effectiveness of learning from the user feedback. In this thesis, only the second area is investigated, which is the effectiveness of feedback learning.

The effectiveness of an RF model depends on the evidence collected based on the feedback from the user. In particular, the user needs to identify as many relevant documents as possible. Based on this evidence, the RF model will identify important index terms that potentially could be used to retrieve more relevant documents for the user. Usually, all the index terms are weighted based on the distribution of the terms in the relevant documents identified by the user. The top ranked terms are selected as the important terms for the user. Normally, the selected terms are used to expand the user query.

However, giving explicit relevance feedback to the system is difficult for the user. It is also time consuming for the user to assess the retrieved documents for relevance. Therefore, there is a need to infer the relevancy of the documents based on measuring the interaction with the user. Such a technique is known as implicit RF (Kelly and Teevan, 2003, White et al., 2005). Examples of implicit measures of relevance are the reading time for the documents, the action of saving or printing the documents and also referencing to the documents (Morita and Shinoda, 1994, Joachims et al., 1997, Konstan et al., 1997, Seo and Zhang, 2000). In the context of browsing, the selection of documents can be used as an implicit indication of relevance, such as selecting the hyperlink while browsing on the Web (Golovchinsky, 1997, Chi et al., 2001, Olston and Chi, 2003, Azman and Ounis, 2004, Vinay et al., 2005) or selecting an image in the *OstensiveBrowser* (Campbell, 2000, Urban et al., 2003).

In this thesis, the effectiveness of an implicit RF system for mobile devices is investigated. An implicit RF system is chosen to overcome two limitations of mobile devices, the small screen size and the limited interaction capability (Vinay et al., 2005). In the next section, issues regarding the implementation of an IR system for mobile

devices are discussed. The motivation for using an implicit RF system for mobile devices is also presented.

2.4 Effective IR system for mobile devices

The advancement of mobile technology makes it possible to access information anywhere and anytime. With *internet-ready* mobile devices, such as mobile phones and PDAs, searching for information becomes much easier. The user may want to use search engines while on the move, such as on the bus or on the train. Thus, the search engine might become one of the most useful applications on mobile devices in the near future.

However, searching for information on mobile devices is problematic due to a number of limitations of the device, such as the small screen size, the limited graphics capabilities, the slower processing speed, the limited input capability and the need for extensive scrolling (Albers and Kim, 2000). A usability study conducted in (Jones et al., 1999) suggested that it is up to 50% less effective to complete a task when the small screen device is used. In the study, it is also suggested that a good system implemented on the small screen device should attempt to provide direct access to information and it should reduce the need for scrolling to find the information (Jones et al., 1999).

The problem of small screen size and the need for scrolling make it difficult to view documents on a mobile device. Chen *et al.* proposed a two-level view for a document, the top view and the bottom view (Chen et al., 2003). The document is divided into several partitions based on the semantic content in the document and each partition is represented by a thumbnail. The thumbnails are displayed in the top level view, representing the overall view for a document. The user may select one of the thumbnails to view the detailed information of the partition in the bottom level view.

The *Power Browser* proposed in (Buyukkokten et al., 2000) provides an efficient Web browsing tool for mobile devices. Similarly, a Web document is partitioned into several *Semantic Textual Units* (STU), where each STU corresponds to a semantic region in a document (Buyukkokten et al., 2001). In order to display all the STUs for the document within the small screen device, a shorter representation of the STU is chosen. A number of representations of STUs are investigated. It is discovered that using the important keywords and the text summary extracted from the STU is the best representation for

the STU (Buyukkokten et al., 2002). Meanwhile, Sweeney and Crestani suggested that a shorter summary is suitable for displaying information from a document on mobile devices (Sweeney and Crestani, 2006). The *SearchMobil* technique annotates the query terms found in the regions in order to assist the user locating the relevant information in the regions (Rodden et al., 2003).

In order to deal with the limitation in processing speed, Lai *et al.* (Lai et al., 2004) proposed the use of a thin-client architecture for displaying a document to reduce the download time and the processing required on mobile devices. The *RSVP* browser is designed to improve navigation capability on mobile devices by providing a rich set of navigational information to the user (de Bruijn et al., 2002). The *information foraging theory* (Pirolli and Card, 1999) is adopted in an automatic image browsing model to find the optimal browsing path for images to be viewed on mobile devices (Liu et al., 2003).

This thesis investigates the effectiveness of using an implicit RF system (Kelly and Teevan, 2003, White et al., 2004) for searching for information on mobile devices. The system attempts to solve the problem of displaying the search results on the small screen size by returning only *four* documents to the user each time. The proposed implicit RF system tackles the problem of limited interaction capability, such as the extensive use of keypad or pen to write a query, by using the implicit RF model to infer the information need of the user from his/her interaction history. Such a system is similar to the *Ostensive Browser* for browsing image collections (Campbell, 2000, Urban et al., 2003).

The implicit RF system proposed in (Vinay et al., 2005) assumes that the information seeking session begins with a query submitted by the user. In response to the query, *four* documents are returned to the user. The user starts browsing through the system by selecting one of the displayed documents and the system responds by returning another set of *four* documents. The browsing continues until the user stops.

However, the overall success of the implicit RF system in satisfying the user's information need depends on the documents he/she selected while browsing. This is because the selected documents are assumed to be relevant by the implicit RF system. The information need of the user is then inferred based on these documents. Therefore, the selection of the documents to be retrieved in the next iteration will depend on the

documents selected in the prior iterations. In other words, the browsing strategy of a user, such as the decision for selecting the displayed documents, determines the overall success of the implicit RF system. Due to the fact that there are a number of ways to choose the documents, the success of a user in his/her information seeking session by using the implicit RF system is not guaranteed. This is apparent as it is shown in the results of the evaluation conducted in Chapter 4 of this thesis. Therefore, there is a need for incorporating a browsing recommendation model to help the user in choosing the documents to ensure a successful browsing session.

2.5 Modelling Browsing Behaviour

Browsing is an important searching strategy in information seeking, especially on the Web. As opposed to query-based searching, the browse-based searching strategy eliminates the need for the user to formulate a query to be submitted to the IR system. Instead, the user moves from one document to another document to find the information he/she is seeking. In general, combining both strategies, query-based and browse-based strategies, is more effective (Olston and Chi, 2003).

Usually, browsing is defined as an action of moving from one document to another document through the hyperlink structure of the Web to find the relevant documents. In this thesis, the browsing definition is generalised to the sequential assessment of documents. For example, the sequential reading of the top retrieved documents (Leuski, 2000), the sequential selection of documents in an interactive RF system and also the browsing through similarity between documents (Smucker and Allan, 2006).

However, browsing is not always an efficient information seeking strategy for the user (Olston and Chi, 2003). One of the reasons is that the user is unfamiliar with the structure that links the documents or the behaviour of the system. For instance, the user may not know the documents that are linked to from the currently visited document, except for the clues given by the anchor text[1] of the hyperlink in the case of Web documents. Moreover, the user is unfamiliar with the behaviour of an interactive IR system, such that he/she is unable to guess the relevance of the documents that will be returned to him/her in the next iteration. Therefore, he/she will not be able to choose a good strategy for browsing and it may lead to an unsuccessful browsing session.

[1] The texts associated with the hyperlink.

The information foraging theory (Pirolli and Card, 1995, Pirolli and Card, 1999) suggests that the user will alter his/her strategies or the structure of the information environment to maximise his/her *rate* of getting valuable information (Pirolli and Card, 1999). Based on this theory, an effective information seeking strategy could be discovered by observing how the actual users interact with the system while looking for information. Therefore, a good browsing recommendation model should be developed based on the browsing behaviour of the user. However, modelling the browsing behaviour of the user is problematic.

2.5.1 Behavioural model for browsing

In (Ellis, 1989), the author observed the behaviour of real users in their information seeking tasks. He suggested that an effective IR system should take into consideration all the behavioural aspects of the user's information seeking strategy in the design of the system (Ellis, 1989). Moreover, it is suggested that the "*berrypicking*" search model is closer to the behaviour of real users in their information seeking tasks (Bates, 1989). The model suggests that real users will start with a broad topic, find relevant documents, refine or modify their information need based on the documents they encounter and will then decide on new directions for their search, which is also known as *evolving search* (Bates, 1989).

The information foraging theory (Pirolli and Card, 1999) indicates that the user will modify the structure of the information environment to maximise the rate of getting the information. The *Scatter/Gather* text clustering browser (Hearst and Pedersen, 1996) was suggested as a solution to provide an optimal information foraging for the user (Pirolli et al., 1996, Pirolli, 1998). The user of the browser will select the best strategy, by choosing the clusters to be expanded or to be combined, such that the experience will maximise his/her rate of getting the relevant information.

It is also suggested that the user is guided by the *information scent* (Chi et al., 2000), which is introduced as part of the information foraging theory. The main assumption is that the user chooses the documents to be selected next based on the *proximal cues*, the snippet or the anchor text of the hyperlink, to justify the content of the linked documents, the *distal page* (Chi et al., 2001). The authors developed two algorithms based on the information scent, the *Web User Flow by Information Scent* (WUFIS) to simulate the potential actions of a user with a given information need, and the *Inferring*

User Need by Information Scent (IUNIS) to predict the user's information need based on the actions of the user (Chi et al., 2001). In addition, the *ScentTrails* method adopts the information scent concept and combines both query-based searching and browsing to guide the user while navigating Web documents (Olston and Chi, 2003).

The technique of inferring information need of the user based on the documents he/she selected while browsing, such as the IUNIS, has been investigated in many occasions (Campbell and van Rijsbergen, 1996, Hirashima et al., 1998, Campbell, 2000, Zhu et al., 2003). It is based on the assumption that the information need of the user is developing as the user browses the documents (Campbell and van Rijsbergen, 1996, Campbell, 2000) and the contents of the documents browsed define the context of the information need of the user (Hirashima et al., 1998). In (Gery, 2002), the potential reading paths are extracted based on the hyperlink structure of the Web and the contents of those documents in the path are used for indexing (Gery, 2002).

Web access logs can also be a source of information to model the browsing behaviour of the user. There are many browsing recommendation techniques that are developed based on Web access logs such as in (Wexelblat and Maes, 1997, Chen et al., 1998, Huberman et al., 1998, Perkowitz and Etzioni, 2000, Mobasher et al., 2001, Mobasher et al., 2002). In (Perkowitz and Etzioni, 2000), the Web access logs are clustered to discover the groups of documents that are frequently accessed together in a browsing session. Each cluster of documents is used to generate an index page, a Web document consisting of the links to the documents in the cluster. Therefore, future users of the Website will benefit since they can quickly reach the documents through the index page (Perkowitz and Etzioni, 2000).

Similarly, Web access logs are clustered to create a set of *usage profiles* or browsing profiles, each profile consists of the documents that are frequently accessed together (Mobasher et al., 2002). The recommendation engine will compute the similarity between the documents already visited by the user in a given browsing state with the clusters of documents, i.e. the usage profiles. The documents in the most similar cluster that have not been visited by the user are then recommended to him/her (Mobasher et al., 2002). Alternatively, the recommendation can be done based on association rules discovered from Web access logs (Chen et al., 1998, Mobasher et al., 2001). Moreover, a map can be generated based on frequently used browsing paths in a Website

discovered from the Web access logs can be used to guide the user in browsing (Wexelblat and Maes, 1997).

The main problem of modelling the browsing behaviour of the user based on Web access logs is that the information about the tasks and the indication of the relevant documents for the tasks are missing in the logs. Therefore, there is no indication whether frequent browsing pattern is a success or a failure in the information seeking session. Most techniques assume that the frequent browsing paths are successful, which could be misleading. Despite of their limitations, Huberman *et al.* (Huberman et al., 1998) used Web access logs to model the browsing behaviour of the user. Their models were based on the assumption that the value of a document is a function of the value of the previous document plus a random variable. A user will only proceed to the next document if the value of the next document is above a given threshold. The model attempts to discover the stopping behaviour of the user while browsing. However, estimating the threshold value for each user for a given task is problematic.

2.5.2 Modelling dependency in browsing

The dependency in browsing specifically refers to the assumption that the value or the relevance of a document may depend on the documents that have been read or assessed previously. For instance, a document could be under-valued if its content is similar to the documents read previously. Similarly, a document has greater value if it consists of novel information that has not been covered by the previously read documents. Therefore, there is a need to model dependency in browsing.

Dependent relevance refers to the concept where the relevance of a document depends on the other documents previously read by the user. On the other hand, *independent relevance* refers to the estimated relevance of a document based on the query submitted by the user independent of other documents. In this thesis, the RSV assigned to the documents by the IR system is the independent relevance value of the documents.

The dependent relevance assumption is not widely used in IR research. However, there are a few approaches that assume that the relevance of a document could depend on other documents. For instance, in the *indirect retrieval method*, the dependent relevance assumption is used to generate a sequence of documents to be returned to the user as a response to his/her query (Croft and van Rijsbergen, 1976). The dependent relevance of

two documents is estimated based on the ratio of the number of common terms in both documents to the number of unique terms in one of the documents. The *cluster-based retrieval* method also assumes that the relevance of a document may depend on other documents (Van Rijsbergen, 1979). It is due to the fact that the retrieval of a document in a cluster depends on the similarity of the cluster's representative (or the centroid of the cluster) to the query of the user and the cluster's representative is computed based on all documents in the cluster. Indirectly, the relevance of a document, in the context of retrieval, depends on other documents in the same cluster.

The dependent relevance assumption has also been investigated in the context of optimal ranking for the retrieved documents. The *Maximal Marginal Relevance* (MMR) ranking method (Carbonell and Goldstein, 1998) assumes that the retrieved documents should be ranked based on their similarity to the user's query and based on their novelty with respect to the documents at higher ranks. The MMR score of a document is a combination of the RSV score of the document and a *novelty* score of the document with respect to the documents at higher ranks. In (Carbonell and Goldstein, 1998), a linear combination is used to compute the MMR score. An alternative combination technique is proposed in (Zhai et al., 2003) as a part of the Risk *Minimization Framework* for IR (Zhai and Lafferty, 2006).

To summarise, the information foraging theory indicates that the user adapts his/her strategy to maximise the *rate* of finding relevant information. Therefore, a browsing recommendation model should be developed based on the browsing behaviour of the user. In this thesis, a browsing recommendation model is developed based on a behavioural model for intertemporal choice (Loewenstein and Prelec, 1992). It is based on the assumption that browsing is an instance of the intertemporal choice problem. The dependent relevance assumption for browsing is also investigated based on the concept of gains and losses in the intertemporal choice model. In the next section, a detailed description of the intertemporal choice model is presented.

2.6 The intertemporal choice model

Intertemporal choice is one of the decision problems discussed in *decision theory*. It is concerned with the preference of individuals for the choices comprising of delayed outcomes or the outcomes that are spread over time. An example of the intertemporal choice problem is the prospect of receiving £10,000 after one year in return to an

investment of £5000 now (Frederick et al., 2002). In this case, the outcome is monetary and the individual is expected to decide whether to give up some amount of money for a bigger return after one year or to keep the existing money.

Intertemporal choice has also been defined as the "*decisions involving tradeoffs among costs and benefits occurring at different point in time*" (Frederick et al., 2002). Moreover, intertemporal choice "*is used to describe any decision that requires trade-off among outcomes that will have their effects at different times*" (Read, 2004). The intertemporal choice problem is modelled by the *discounted utility model* (DU) (Samuelson, 1937).

It is also important to note that there are two types of models in decision theory, the *normative model* and the *descriptive model* (Kahneman and Tversky, 1979). On the one hand, the normative model identifies the best decision by assuming that people are fully aware of the consequences of the choices and are fully rational. Therefore, the model will suggest the best decision for an individual based on how he/she *should* make the decision. On the other hand, it often happens that people do not behave in an optimal way when making decisions. For instance, in the case of the risky choice, people may prefer a smaller reward with certainty rather than a riskier and much bigger expected reward. Therefore, an alternative model, the descriptive model for the decision problem, is developed to model the *actual* behaviour of individuals in making a decision. Usually, the descriptive model is developed as the solution to tackle the *anomalies* of the normative model. The anomalies of the normative model refer to a situation where the preference of an individual does not follow the one chosen by the normative model. The anomalies of the discounted utility model for the intertemporal choice problem will be discussed in detail later in this section.

In the next section, the discounted utility model (Samuelson, 1937) for the intertemporal choice problem as well as its anomalies are presented.

2.6.1 The discounted utility model (DU)

According to Frederick *et al.* (Frederick et al., 2002), the topic of intertemporal choice was established as early as 1834 with the publication of *The Sociological Theory of Capital* by John Rae. Since then, many researchers viewed the psychological motives of *time preference*, the preference for immediate *utility* over a delayed *utility* (Frederick

et al., 2002), as a blend of different intertemporal motives. A significant development of the problem was achieved when DU was proposed by Samuelson in (Samuelson, 1937) where those motives that has been extensively discussed for a century were compressed into a single parameter, the discount rate (Frederick et al., 2002). A good discussion on the historical development of DU is given by Frederick *et al.* in (Frederick et al., 2002).

Since the introduction of DU by Samuelson in 1937, it has dominated the economic analyses of the intertemporal choice problem. For two given sequences of *consumption levels*, (i) $(c_0,...,c_T)$ and (ii) $(c'_0,...,c'_T)$, the sequence (i) is preferred to the sequence (ii) if and only if,

$$\sum_{t=0}^{T} \delta^t \times u(c_t) > \sum_{t=0}^{T} \delta^t \times u(c'_t), \tag{2-1}$$

where δ is the discount factor for one period and $u(c_t)$ is the *utility* or the *subjective value* for the consumption at time t given by a concave utility function (Loewenstein and Prelec, 1992). In essence, the *overall utility* of a sequence of consumption levels is the *discounted* sum of the utilities of the consumption levels and the sequence that yields a better overall utility is preferred.

There are *seven* characteristics of DU as presented in (Frederick et al., 2002). First, DU assumes that the individuals assess new strategies by integrating them with their existing strategies. For instance, an individual may have a budget plan for his/her incomes and expenses for a few years to come. He/she is offered an option of investing some of his/her money now for a bigger amount of money in return after a year. His/her new budget plan will include a reduction of some money he/she invests now and an increment of money he/she will gain after a year. Therefore, the intertemporal choice problem is whether to choose the old budget plan or the new budget plan.

Second, DU makes an assumption that the overall utility of a sequence of outcomes remains unchanged over time. This means that the subjective value perceived by the individuals for the sequence of outcomes at a given time will be the same if it is viewed at later times. This *utility independence* (Frederick et al., 2002) assumption makes it easy to model the intertemporal choice problem, since the distribution of the overall

utility across times is neglected. For instance, in the earlier example, the overall utility of an individual on the new budget plan will be the same even if it is re-assessed after a few months.

Third, DU assumes that the utility of any outcome in the sequence is not affected by the utility of previous or future outcomes that might have or will be experienced. For instance, the utility of getting some amount of money will not be affected by the potential of losing some money in the future. This scenario is referred to as the *consumption independence* property in (Frederick et al., 2002). Fourth, DU also assumes that the utility of any given outcome in the sequence as perceived by the individuals will be consistent over times. For instance, the subjective value of the individuals for getting a given amount of money will be the same if it is evaluated now or later. This property is called the *stationary instantaneous utility* in (Frederick et al., 2002).

Fifth, DU assumes that there exists a common discount rate for all types of outcomes of the individuals. It means that, the preference (or decision) behaviour of the individuals is consistently independent of the outcome type. This assumption eliminates the need for having a unique discount rate for each type of outcome, such as a different discount rate for money-related outcomes and for health-related outcomes. Sixth, DU assumes a constant discounting where the discount rate for each period is the same. Finally, the seventh property of DU is that it assumes a *positive time preference* and a *diminishing marginal utility* (Frederick et al., 2002), in which an immediate utility is preferred over a delayed utility.

In the next section, a detailed discussion of the anomalies of the DU is presented.

2.6.2 The anomalies of DU

Samuelson made no claim on the type of decision model for DU, whether it is a normative model or a descriptive model for the intertemporal choice problem (Frederick et al., 2002). Despite his reservations, the model has been adopted as the 'standard' framework for analysing the intertemporal choice problem (Frederick et al., 2002). However, the findings from empirical studies on intertemporal choice have suggested various insufficiencies of the DU model as a descriptive model for the intertemporal

choice problem. The various shortcomings of the model are referred to as the anomalies of the DU model.

It has been discovered that *hyperbolic discounting* better describes the decision behaviour of individuals in the intertemporal choice problem as compared to *exponential discounting* (Loewenstein and Prelec, 1992, Read, 2004). Exponential discounting refers to a constant rate of time preference, while the hyperbolic discounting refers to the declining rate of time preference (Frederick et al., 2002). This phenomenon is observed in empirical studies conducted by a number of researchers (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Frederick et al., 2002, Lazaro et al., 2002). Frederick *et al.* (Frederick et al., 2002) described four pieces of evidence for hyperbolic discounting.

The first evidence is based on the analysis of the behaviour of individuals in comparing a smaller-sooner reward to a larger-later reward (Frederick et al., 2002). It can be illustrated based on the example by Richard Thaler in 1981 (Loewenstein and Prelec, 1992, Frederick et al., 2002, Read, 2004). The *subjects* were asked to indicate the amount of money that will be needed in a month, a year and ten years, such that the utility of receiving them at those times is equivalent to getting $15 now. Based on the median taken from the responses, the amount of money they need for delaying receiving $15 now is $20 for a month, $50 for a year and $100 for ten years. It implies a discount rate of 345% for a month period, 120% for a year period and 19% for a ten years period (Frederick et al., 2002). Therefore, a declining discount rate is observed. Similar results were discovered in (Benzion et al., 1989, Green and Myerson, 1996).

The second evidence is based the observation that the mathematical hyperbolic functions nicely fit the empirical data (Rachlin et al., 1991, Loewenstein and Prelec, 1992, Myerson and Green, 1995, Cairns and van der Pol, 2000, Frederick et al., 2002, Lazaro et al., 2002). The third evidence is based the *preference reversal* observed in empirical studies. For instance, an individual may prefer $110 in 31 days over $100 in 30 days, but also favour $100 now over $110 tomorrow (Frederick et al., 2002). The final evidence is based on the analysis conducted in (Frederick et al., 2002) on empirical data compiled from many studies. Based on the regression analysis, they discovered that the discount rate declines (Frederick et al., 2002).

The hyperbolic discounting anomaly is probably the most investigated phenomenon. There are several other anomalies discovered for the DU model. The first anomaly is known as the *sign effect* (Frederick et al., 2002, Read, 2004) or the *gain-loss asymmetry* (Loewenstein and Prelec, 1992). It is concerned with the behaviour of individuals in which the gains (or the positive outcomes) are discounted at higher rates as compared to the losses (or the negative outcomes). Based on a study conducted by Loewenstein, the individuals are indifferent about getting $10 now and $21 in a year, but are also indifferent about losing $10 now and $21 in a year.

The second anomaly is known as the *magnitude effect* (Loewenstein and Prelec, 1992, Frederick et al., 2002, Read, 2004), where smaller outcomes are discounted more than larger outcomes (Frederick et al., 2002). Based on the experiment conducted by Richard Thaler in 1981, which was widely discussed by many researchers, the individuals who were indifferent between getting an immediate $15 and $60 in a year, were also indifferent between an immediate $250 and $350 in a year, as well as between $3000 now and $4000 later in a year (Loewenstein and Prelec, 1992, Frederick et al., 2002, Read, 2004).

The third anomaly is known as the "*delay-speedup asymmetry*" (Loewenstein and Prelec, 1992, Frederick et al., 2002), where the amount to compensate a delay for an outcome that is supposed to be received immediately is higher than the amount to compensate for a speed-up of an outcome that is supposed to be received later. In the example in (Frederick et al., 2002), the individuals are willing to pay $54 to speed-up the delivery of a VCR for a year (to be received immediately) but will demand $126 for a year delay of the VCR that is supposed be received now.

The fourth anomaly is the "*preference for improving sequence*", where many researchers have found that for some intertemporal choice problems, *negative time preference* is more desirable, such as the preference for improving salary and health (Loewenstein and Prelec, 1991, Loewenstein and Prelec, 1993, van der Pol and Cairns, 2000). The fifth anomaly is the violation of the independence of the consumption, where the utility of a given outcome may depend on the outcomes experienced previously or in the future, such that it violates the assumption of consumption independence of the DU model.

Based on the anomalies discussed, a descriptive model for the intertemporal choice problem is required such as the behavioural model for intertemporal choice is as proposed by Loewenstein and Prelec (Loewenstein and Prelec, 1992). A detailed discussion of the model will be presented in Section 3.2. In this thesis, the browsing problem is treated as an intertemporal choice problem. The behavioural model for intertemporal choice proposed in (Loewenstein and Prelec, 1992) is adapted to model the browsing behaviour of the user of an IR system. The proposed behavioural model for browsing is called the *intertemporal choice browsing model* and it will be discussed in the next chapter.

2.7 Chapter Summary

In this chapter, the general model for an IR system was presented (Section 2.2). The role of the user to improve the effectiveness of an IR system was also discussed, particularly in the context of RF system (Section 2.3). The issues of implementing an IR system on mobile devices were reviewed and the motivation for using an implicit RF system for mobile devices is established (Section 2.4). In this chapter, some of the approaches for modelling browsing behaviour of the user are reviewed (Section 2.5). Finally, a detailed discussion on the intertemporal choice model was given (Section 2.6).

3 The Intertemporal Choice Browsing Model

3.1 Introduction

In this thesis, it is argued that the effectiveness of an information retrieval (IR) system depends on the ability of the user to make the right decision at every step of browsing in his/her information seeking task. It includes the selection of keywords for a query to represent his/her information need or the selection of documents to be read in order to find the information he/she is looking for. Due to the fact that the user usually is unfamiliar with the nature of the document collection and the retrieval model adopted by the IR system, the capability of the user to make the right decision is often limited. Therefore, it is essential for the system to provide more assistance to the user in his/her information seeking session.

This thesis focuses its investigation on the problem of browsing in IR. In particular, the aim of this thesis is to propose an effective browsing recommendation model for IR. In this thesis, an *intertemporal choice model* (Loewenstein and Prelec, 1992, Frederick et al., 2002, Read, 2004) is adopted to model browsing behaviour of users. It is based on the assumption that browsing is an example of the intertemporal choice problem. Therefore, the intertemporal choice model is used to model user decision making in browsing. Based on the model, a good browsing strategy could be recommended to the users.

In this chapter, a behavioural model for intertemporal choice is described (Loewenstein and Prelec, 1992). The model has been extensively studied in economic and social studies (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Frederick et al., 2002, Lazaro et al., 2002) and the model is developed based on the behaviour of people dealing with the intertemporal choice problem (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Frederick et al., 2002, Lazaro et al., 2002). Due to the success of the model in economics, it is hoped that the decision behaviour of users in browsing can be modelled by the intertemporal choice model.

In Section 3.2, a detailed description of the behavioural model for intertemporal choice is presented. In Section 3.3, the implementation issues of applying the intertemporal choice model for browsing are discussed. A detailed discussion on the contribution of the model into IR research is presented in Section 3.4. Finally, a brief summary of this chapter is provided in Section 3.5.

3.2 The Intertemporal Choice

The problem of intertemporal choice was discussed in Section 2.6. The discussion focused on the application of the model in the economics context. By definition, the intertemporal choice is "*used to describe any decision that requires trade-offs among outcomes that will have their effects at different times*" (Read, 2004). In this chapter, the use of the model is extended to modelling browsing behaviour of the users in IR.

3.2.1 The Behavioural Model for Intertemporal Choice

First, the behavioural model for intertemporal choice proposed in (Loewenstein and Prelec, 1992) is described in this section. Let $X = \{x_1, t_1; ...; x_n, t_n\}$ be a set of outcomes, where x_i is the outcome delayed for t_i period of time. t_i can also be viewed as the time for receiving the outcome. The model defines the utility of the outcomes as:

$$U(X) = U(x_1, t_1; ...; x_n, t_n) = \sum_{i=1}^{n} v(x_i)\phi(t_i) \qquad (3\text{-}1)$$

where $v(x)$ is the *value function* for an outcome x, $\phi(t)$ is the *discount function* for the delay t and $t_n > t_1$. Let $X' = \{x_1', t_1; ...; x_n', t_n\}$ be another set of outcomes. According to the model, the sequence of outcomes X will be preferable to the sequence of outcomes X' if and only if $U(X) > U(X')$ and both sequences of outcomes are equivalent

when $U(X) = U(X')$. $U(X)$ and $U(X')$ are the utility or subjective value of X and X', respectively.

In essence, the model consists of two components, the value function and the discount function. The value function is applied to the outcome in order to transform the value of the outcome to a subjective value. The subjective value of the outcome is also referred to as the *instantaneous utility* of the outcome (Frederick et al., 2002). The discount rate for each delayed outcome is modelled by the discount function. The value function and the discount function will be discussed in Section 3.2.1.1 and in Section 3.2.1.2, respectively.

3.2.1.1 Value function

In order for the model to be descriptive, two important characteristics of the value function need to be discussed, loss aversion and the sensitivity towards the value of the outcome. These characteristics are discussed in this section.

Based on many studies conducted in the area of economy, people tend to overestimate the value of losing and underestimate the value of gaining (Kahneman and Tversky, 1979, Loewenstein and Prelec, 1992). Thus, the role of the value function is to transform the objective value of an outcome into a subjective value to capture this behaviour. Such a function is normally a monotonic increasing concave function for positive outcomes and a monotonic decreasing convex function for negative outcomes (Kahneman and Tversky, 1979).

The second characteristic of the value function is its sensitivity towards the value of the outcome. It is related to the scenario whereby different values for an outcome may be perceived depending on the situation. In this context, the Prospect Theory as well as the behavioural model for intertemporal choice suggested that the value of the outcome should be estimated with respect to the total *wealth* of an individual or the *gains/losses* suffered by an individual (Kahneman and Tversky, 1979, Loewenstein and Prelec, 1992). This characteristic is mainly important in economic applications, but the concept is general.

For instance, an individual who has in his/her possession £1000 is about to either receive or lose £50 in a bet. In this view, an individual will end up with either £1050 or

£950 in the final state of the wealth after the bet. Therefore, in the case of the total wealth, an individual will perceive that the outcome of the bet is either having £1050 or £950 in the end. In another view, an individual may end up gaining £50 or losing £50. Therefore, in the case of the gains or losses, an individual will perceive that the outcome of the bet is either getting an extra £50 or losing £50. It means that, the choice of an individual in the bet is either to gain some money or to lose some money without taking into consideration his/her status quo (current money that he/she has).

Assuming that the outcomes are viewed as gains and losses, another property of the value function should also be discussed. According to the Prospect Theory, which is adopted by the Loewenstein and Prelec model, the value function for losses is steeper than the value function for gains (Kahneman and Tversky, 1979). This is to accommodate the loss aversion behaviour of an individual in decision making with respect to gains and losses. For instance, most people feel that the subjective value for losing £50 exceeds the subjective value for gaining £50. They overestimate the value of losing as compared to the value of gaining the same amount of money. Therefore, the value function is steeper for losses than for gains.

Based on the characteristics discussed, a number of mathematical functions can be used to model the value function. In the literature, there is no specific function claimed to perform the best. Therefore, in this thesis the following *log* function is used to model the value function for gains (or for positive value) and for losses (or for negative value).

a. Gains (positive outcomes)

$$v(x) = \log_{10}(ax+1) \tag{3-2}$$

where $x \geq 0$ and $a > 0$.

b. Losses (negative outcomes)

$$v(x) = -\log_{10}(-bx+1) \tag{3-3}$$

where $x < 0$ and $b > 0$.

Variable a controls the rate of change for the gains function and variable b controls the rate of change for the losses function. Since the losses function is steeper than the gains function, b is always greater than a.

In the context of information retrieval (IR), each outcome is a document. Each value of the outcome is the value of a document with respect to the user's information need. Usually, an IR model will assign a value to each document before retrieving it, such as the retrieval status value (RSV). The RSV value for a document is often estimated based on the query submitted by a user. However, the meaning of the RSV value is often based on the assumption used in a particular IR model.

The challenge is to interpret the value of document(s) in term of *wealth*, *gains* and *losses*. The issues will be elaborated in Section 3.3.3.

3.2.1.2 Discount function

The second component of the model is the discount function. It replaces the probabilistic aspect of the outcome in the expected utility model (von Neumann and Morgenstern, 1944) with a decreasing weighting function with respect to the time of the outcome. Over the years, researchers have focused on deriving a good discount function based on the empirical data (Loewenstein and Prelec, 1992, Loewenstein and Prelec, 1993, Cairns and van der Pol, 2000, Lazaro et al., 2002).

There are two properties of the discount function, the type of discounting and the shape of the discount function. *Time preference* refers to the preference of an individual for immediate utility or delayed utility (Frederick et al., 2002). *Positive time preference* is used to describe the preference of an individual for immediate utility over delayed utility and *negative time preference* is used to describe the preference of an individual for delayed utility over immediate utility (Loewenstein and Prelec, 1991, Loewenstein and Prelec, 1993, van der Pol and Cairns, 2000, Frederick et al., 2002).

For the positive time preference, a monotonic decreasing function with respect to the delay is used to model the discount rate for each delayed outcome, or a positive time discounting. In this case, the delayed outcomes are discounted more than the immediate outcomes. The majority of the empirical studies conducted for the intertemporal choice

problem are based on the positive time preference (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Lazaro et al., 2002). The negative time preference is treated as an anomaly of the DU, where in some cases individual prefers an increasing utility over time such as an increasing salary, an increasing quality of life or an improvement of health (Loewenstein and Prelec, 1993). The negative time preference is modelled by a monotonic increasing function with respect to the delay, or a negative time discounting.

Loewenstein and Prelec defined two different motives for the two types of discounting as *impatience* and *preference for improvement* (Loewenstein and Prelec, 1993). In the case of impatience motive, people prefer a positive time preference. The negative time preference is more appropriate for the preference for improvement motive.

There are two shapes of discount function, an *exponential* discount function and a *hyperbolic* discount function. In (Samuelson, 1937), Samuelson suggested that the discount rate for each period of time is constant such that the discount factor in each period is the same. He proposed an exponential discount function to model the constant discount rate for all period. The discount function proposed by Samuelson is as follows (Samuelson, 1937):

$$\phi(t) = \frac{1}{(1+r)^t} \tag{3-4}$$

where t is the delay and variable r controls the discount rate of the function.

However, evidence from recent empirical studies suggested that an exponential discount function does not fit a lot of empirical data (Loewenstein and Prelec, 1992, Green and Myerson, 1996, Cairns and van der Pol, 2000). Based on the data, the discount rate for each period declines as the delay increases. Therefore, a number of hyperbolic functions have been proposed for the discount function (Loewenstein and Prelec, 1992, Frederick et al., 2002, Read, 2004). A more detailed discussion of hyperbolic discount functions was given in Section 2.6.2.

Most of the discount functions proposed is parametric. The parameter(s) of the functions are used to make the function more flexible such that it will fit nicely to

model the empirical data. In this thesis, two widely used single-parameter functions proposed by Mazur (Read, 2004) and Harvey (Harvey, 1986) are chosen for the intertemporal choice model. The functions with single parameter are chosen because it will be easier to tune its performance. The following are the hyperbolic functions used in this thesis:

a. Harvey (Harvey, 1986)

$$\phi(t) = \frac{1}{(1+t)^h} \tag{3-5}$$

b. Mazur (Read, 2004)

$$\phi(t) = \frac{1}{(1+st)} \tag{3-6}$$

where t is the delay, while variable h and s control the discount rate of the functions.

The behavioural model for intertemporal choice proposed in (Loewenstein and Prelec, 1992) is presented to resolve the anomalies of the discounted utility (DU) model described in Section 2.6.2. In this thesis, the model is adopted to model the browsing behaviour of the user. It is based on the assumption that browsing is an intertemporal choice problem. In the next section, the intertemporal choice model for browsing is described. In this thesis, the model is known as the *intertemporal choice browsing model* for IR.

3.3 Browsing as an Intertemporal Choice Problem

Browsing is an important strategy for information seeking (Bates, 1989). Browsing is guided by the information need of the user (Chi et al., 2001), where the next document to be chosen is decided upon based on the anticipation that the document contains the information required by the user. Therefore, the success of a browsing session depends on the documents selected by the user while browsing. Moreover, the objective of browsing is to find the relevant documents and to find them as quickly as possible.

In order to make an optimal decision while browsing, the user should be fully informed about the consequences of all possible strategies. However, the user does not have enough knowledge about the document collection or the behaviour of the system for him/her to be fully informed. He/she has no knowledge about the documents in the

collection and how they link with each other in the case of the Web. Even with the title, document summary or snippet displayed by the search engine on the search result page, the actual content of the document remains hidden from the user unless the user has prior knowledge of the document. Therefore, it is unlikely that he/she will be able to anticipate the outcomes of all the strategies available for him/her. Moreover, it is believed that a user would not be able to make an optimal decision for his/her browsing strategy.

Due to the user's inability to make an optimal decision, his/her browsing session may be sub-optimal or unsuccessful. Meanwhile, the system has the knowledge about the documents in the collection based on its indexes. By anticipating all possible user actions, such as selecting a document as an implicit indication of relevance in the RF system, the system would be aware of all possible responses to those actions. Therefore, with such knowledge, the system is fully informed and it is capable of providing assistance to the user in making a decision on the strategy for browsing.

If the system is capable of anticipating all possible browsing strategies of a user for a given query and is fully informed of the consequences of the strategies, the system can recommend the optimal browsing strategy to the user. Therefore, the task of the system is to discover all the possible browsing strategies, to evaluate the consequence of each strategy to the user and to recommend the strategy that will result in the best outcome of the browsing session for the user. The system could assign a score to each strategy indicating how that strategy would benefit the user and it should recommend the strategy with the highest score.

The problem involving the decision to choose the best strategy for browsing fits the model for intertemporal choice. The problem of intertemporal choice is to choose among a set of strategies where each strategy consists of a set of outcomes received at different times. In the case of browsing, each possible browsing path consists of a set of documents viewed at different times or in sequential order. Therefore, the decision problem for browsing can be viewed as an intertemporal choice problem.

Ideally, each browsing path can be treated as a strategy and all documents in the browsing path can be treated as the outcomes. The problem of assigning a score to a browsing path is similar to the problem of intertemporal choice. Assuming that each document has a value, such as the RSV value computed for retrieval, and a browsing path consists of a sequence of documents, the intertemporal choice model can be used to compute the utility of a browsing path. Based on this assumption, each document is an outcome and each browsing path is a set of outcomes that a user receives, each at different time.

In the next section, the decision strategy to select the recommended browsing path is presented. It is based on the idea that the utility for each browsing path can be computed by using the intertemporal choice model. Furthermore, various issues regarding the value estimation problem for the documents and the discount functions for browsing are also discussed.

3.3.1 Browsing path recommendation

A browsing session can be viewed as a series of states in the information seeking process, where each state occurs at a specific time. In a typical browsing scenario, the states are represented by the documents assessed by the user, one at a time. In this case, the browsing session is a sequence of documents. The states can also be represented by the set of documents displayed by the system as a result of the user's interaction. For instance, the states can be represented by the top retrieved documents displayed by an implicit relevance feedback (RF) system iteratively (Campbell, 2000, White et al., 2004) as a result of the user's feedback, such as choosing one of the displayed documents. In this case, a sequence of system responses (which are displaying a set of documents) is generated by the series of interaction from the user. Therefore, a state can be represented by a set of documents at each time. Without loss of generality, each state is represented by a single document at this point of time.

A browsing path can be defined as a sequence of documents read by the user in one session. Let D be a set of documents in a collection and R be a subset of the documents where $R \subseteq D$ and $|R| = n$. A browsing path is defined as an ordering of $d_i \in R$ such that $path_k = \{d_1, t_1; ...; d_n, t_n\}$, where t_i is the time of reading (or the position of) document

d_i in the browsing path $path_k$ and $2 \le |path_k| \le n$, where a browsing path consists of at least two documents. Let $path$ be the set of all possible browsing paths derived from R and $path_k \in path$. In order to simplify the discussion, the length of each browsing path is assumed to be n at this moment.

Let U_k be the utility or the score for $path_k$. $path_k$ is chosen as the best browsing path (the most preferred browsing path) if and only if U_k is the highest utility, such that:

$$U_k = \max\{U_1,...,U_n\} \tag{3-7}$$

where U_i is the utility of $path_i$. Note that there is no restriction on the number of documents in each state of a browsing path as long as a single value can be derived from the documents in the sequence. For instance, each sequence in the browsing path can consists of more than one document and an average value of the documents can be used for the sequence.

The set of browsing paths can be generated by considering all possible browsing strategies of a user for a given query. In the case of the Web, each browsing path consists of the documents that can be browsed by the user through the hyperlink structure. In the case of an interactive IR system, each browsing path consists of a sequence of documents assessed by the user through his/her interaction with the system. Moreover, the set of browsing paths can consist of all possible sequences of documents selected and read from the top-ranked retrieved documents.

3.3.2 The utility of a browsing path

Based on the recommendation strategy described in the previous section (Section 3.3.1), a technique to measure the utility of a browsing path is required. The utility (or the subjective value), U_k, for a browsing path, $path_k$, is measured by using the intertemporal choice model described in Section 3.2.

First, let us assume that a document has a single value and the value is certain. Let u_i be the value of document d_i at time t_i. In this thesis, the notation u is used to represent the value of a document and the notation x is used to represent the value of an outcome in general context. The sequence of documents in a browsing path $path_k$ can be represented by $U(u_1,t_1;...;u_n,t_n)$. Therefore, based on Equation (3-1), the utility for browsing path $path_k$ is estimated by:

$$U_k = U(u_1,t_1;...;u_n,t_n) = \sum_{i=1}^{n} v(u_i)\phi(t_i) \tag{3-8}$$

The subjective value $v(u_i)$ can be computed by Equation (3-2) or (3-3) depending whether $u_i \geq 0$ or $u_i < 0$, respectively. The discount function $\phi(t_i)$ can be computed by using either Equation (3-4), (3-5) or (3-6).

In (Wheeldon and Levene, 2003), a similar technique is proposed to rank trails for in a website. Each trail consists of a sequence of Web documents connected by hyperlinks. The algorithm discovers and ranks a set of trails from a defined starting page, which is very similar to the model proposed in the chapter. Moreover, the score of each trail is computed based on the scores of the documents in the trail. In addition, a discount function is incorporated in two of the three trail scoring techniques proposed.

The model proposed in this thesis is, however, based on the assumption that browsing is an intertemporal choice problem. Accordingly, the behaviour of browsing could be modelled by the intertemporal choice model. Therefore, the results of the extensive studies conducted on the intertemporal choice problem in other areas can directly be adopted, such as the choices of discount functions and value functions to be used. It is hoped that browsing behaviour of the user can be modelled through this investigation.

3.3.3 Value function

As discussed in details in Section 3.2.1.1, there are two issues related to the value function, the loss aversion behaviour and the sensitivity of an individual towards the value of the outcome. The loss aversion can be dealt with by transforming an objective value of an outcome to a subjective value by applying a monotonic increasing concave function or a monotonic decreasing convex function to the value. Moreover, people are

more sensitive to gains/losses than total wealth and the value function for losses is steeper than the value function for gains.

Based on Equation (3-2) or (3-3), the goal of a value function is to compute a subjective value from an objective value, x_i, or u_i in the case of the value of a document. In browsing, u_i should reflect the objective value of a document at time t_i with respect to the user's information need. The function $v(u_i)$ is a monotonic increasing concave function for positive u_i or a monotonic decreasing convex function for negative u_i to deal with the loss aversion. However, the estimation of u_i for a document remains one of the main challenges of the model.

The terms wealth and gains/losses are more meaningful in the economic domain. In the case of IR, the *collective value* and *differential value* are used to describe wealth and gains/losses, respectively. In the following sections, those terms elaborated in the context of IR.

3.3.3.1 Collective value (Wealth)

In the economic domain, wealth is often associated with the total amount of money or properties owned by an individual at a given time. Let $x_1, x_2, ..., x_n$ be a set of monetary outcomes for time $t_1, t_2, ..., t_n$. The total wealth of the individual at time t_n is:

$$\sum_{i=0}^{n} x_i \tag{3-9}$$

Equation (3-9) is valid since the outcome x is additive. In general, the value of documents or information is not additive (Rafaeli and Raban, 2003). The value of reading two pieces of information may not be the same as the summation of both values. The total value of document u_1 and u_2 may not be equal to the value of $u_1 + u_2$. Equation (3-9) is not meaningful since the value of documents may not be additive. Therefore, the concept of wealth is redefined as the collective value in IR domain.

The collective value *refers to the value of a set of documents or a set of information with respect to a given information need or a given query when they are assessed collectively*. It is the value of a set of documents perceived by the user when he/she

reads all the documents in the set. There are two ways to view the collective value, the value of an *aggregate document* against a query or the value of an *aggregate query* against a document.

The aggregated document is a set of documents combined together to create a single surrogate document. The aggregate document encompasses the contents of all documents in the set. For instance, the cluster representative (or the centroid) of a cluster of documents can be used as an aggregate document (Van Rijsbergen, 1979). The collective value of the set (cluster) of documents is the value of the aggregate document against a given query. In the case of browsing, the aggregate document can be derived from the documents read so far including the next document in the browsing path. Therefore, the collective value of the browsing path is the value of a given query against the aggregate document.

An aggregate query is a dynamic query derived from a set of documents read by the user. It is assumed that the documents already read by the user define the context of his/her information need (Hirashima et al., 1998). Thus, the query is refined, based on the documents already read by the user in the browsing session. One example use of the aggregate query is the Ostensive Model (Campbell and van Rijsbergen, 1996). In the model, the information need is developed as the user reads more documents. Therefore, a query at a given stage of browsing is computed dynamically based on the documents already visited by the user. The collective value is the value of a document based on the aggregate query.

For instance, let us assume that a user has read documents d_1, d_2 and d_3. The user is about to choose document d_4, d_5 or d_6. The possible browsing paths of the user are $p_1 = (d_1, d_2, d_3, d_4,)$, $p_2 = (d_1, d_2, d_3, d_5)$ or $p_3 = (d_1, d_2, d_3, d_6)$. In the conventional method, the *independent relevance* assumption, the value of document d_4 is estimated based on a given query q, such as $u_4 = rsv(q, d_4)$. In the case of the collective value, the value of document d_4 is estimated together with the other documents, d_1, d_2 and d_3. Let us assume that d_{1234} is an aggregate document for d_1, d_2, d_3 and d_4. The collective value for document d_4 is estimated based on measuring the retrieval status value (RSV) for the aggregate document d_{1234} against the query q such that $u_4 = rsv(q, d_{1234})$.

Assuming that d_1, d_2, d_3 and d_4 are a member of a cluster in a cluster-based retrieval model (Van Rijsbergen, 1979) and d_{1234} is the centroid of the cluster, the RSV of document d_4 is equal to the RSV of the centroid of the cluster. Meanwhile, the aggregate query is the *refined* query, q_{123}, generated after documents d_1, d_2 and d_3 have been read, such as through multiple RF iterations. The collective value for d_4 is estimated by the following equation, $u_4 = rsv(q_{123}, d_4)$.

3.3.3.2 Differential value (gains or losses)

Similarly, gain and loss are modelled as differential value in the context of browsing. In monetary system, a *gain* or a *loss* can simply be modelled by whether a new outcome will increase (gain) or decrease (loss) the total wealth of an individual as compared to his/her status quo (such as the example in Section 3.2.1.1). However, the model cannot be applied in the context of browsing since the value of documents or information is not additive. Therefore, the concept of gains/losses can be redefined as the differential value.

The differential value *refers to the value of a document perceived by the user with respect to prior documents already read in a given browsing session.* The differential value of a document depends on the documents already read. For instance, a document is undervalued if a similar document has been previously read. Similarly, a document is overvalued if it contains novel information to the user. Based on this scenario, the differential value of a document should be higher when it delivers novel information and it should be lower when the document does not give novel information.

Novelty models (Zhang et al., 2002, Allan et al., 2003, Gabrilovich et al., 2004) can be used to estimate the differential value of a document. The novelty score used in these models measure the novel information contained in a document with respect to other documents. In the context of browsing, the novelty score of a document depends on the new information in a document that has not been covered in the previous documents. The novelty score alone may not be sufficient to act as the score for a document since the new information may not be useful to the user if it does not satisfy his/her information need. Therefore, the novelty score of a document can be combined with its relevance score as a single differential value for a document. Such combination has

been explored by Carbonell and Goldstein in the Maximal Marginal Relevance (MMR) ranking technique (Carbonell and Goldstein, 1998) and Zhai *et al.* in (Zhai et al., 2003).

3.3.3.3 Independent value

Both Section 3.3.3.1 and Section 3.3.3.2 are similar in the way that the value estimated for a document depends on other documents (previously read) by computing its collective or differential value. Such a model is known as a dependent relevance model, where the relevance or the value of a document depends on the other documents. Most IR models do not incorporate the dependent relevance assumption, because it is more complicated. Therefore, a value of a document is assumed to be independent of other documents and is usually computed against the query statement of a user.

Even though the dependent relevance assumption is not so popular in IR literature, it is believed to be important in the context of browsing. When a user sequentially reads the documents while browsing, the quality of the next document to be read with respect to a particular topic will depend on how much a user has read about the topic in the previous documents, at least in the same browsing session.

The collective value and the differential value are important due to the fact that they are inspired by the concept of *wealth* and *gains/losses* in the economic domain. People are more sensitive to those values than to the actual value of an outcome. In IR, there is a potential that the user is sensitive to the dependent relevance measure, especially in the context of browsing.

Meanwhile, most IR models assume that the value of a document depends only on the query of the user, such as the probability of relevance of a document for a given query in the probabilistic model for IR or the distance measure between the query vector and the document vector in Vector Space Model. Since this assumption is well accepted in IR community and there are many models developed based on this assumption, it is reasonable to use the independent value of the document in the intertemporal choice model for browsing. The independent value *refers to the value of a document assessed against the information need or the query statement.* Most retrieval models can be used to estimate the independent value, which is often called the retrieval status value (RSV) or the relevance value of a document.

3.3.3.4 Uncertain outcomes

Due to the fact that the intertemporal choice model was proposed to deal with decision problems in economy, the outcome is often interpreted as a monetary outcome. The value of such outcome is usually certain. For instance, a note of £100 will be valued as one hundred pound sterling and the value will be the same for all people. However, when the outcome is a document or information, its value may not be certain. The value is normally estimated based on the information need or the query. The relevance of a document is dynamic, such that the relevance of a document changes over time, and multi-dimensional, such that the relevance of a document is different for different users (Borlund, 2003), which leads to a number of different values that could be assigned to the document.

The issue of uncertainty in the value of the document may not affect the conventional retrieval system since most systems rank the retrieved documents. As long as it follows the Probabilistic Ranking Principle (PRP) (Robertson, 1977) by ranking documents in decreasing order of their relevance to the user's information need, it will maximise the expectation of finding the relevant documents (Gordon and Lenk, 1991). Only the relative value of documents not the actual values are typically important to rank the documents.

On the other hand, the actual value (if possible) is very important in the case of the intertemporal choice model and an extension of the proposed model discussed in Section 3.2 is required to deal with this problem of uncertainty in the value of documents. In order to deal with uncertain outcomes, the intertemporal choice model should be general enough to incorporate uncertainty of the outcomes. Instead of assuming an outcome with a certain value, the model should consider an outcome with a set of weighted values or a set of weighted outcomes for a given time. Without loss of generality, it is assumed that the weight is probabilistic in nature where each weight is between 0 and 1 and the total weight for the set of outcomes or for the set of values for an outcome is equal to 1. Indeed, the weight does not have to be a probability measure. A simple normalization technique can be applied to those weights that do not satisfy the condition.

Referring to Equation (3-1) in Section 3.2.1, x_i is the value of an outcome at time t_i. In the context of browsing, u_i is the value of a document d_i at time t_i. Now, let the

document consists of a set of possible values and a set of weights, one for each value, such that $u_i = \{u_{i1}, w_{i1}; ...; u_{im}, w_{im}\}$, where w_{ij} is the weight of the document to have the value of u_{ij}. The subjective value of the document, u_i, is given by:

$$v(u_i) = E(u_{i1}, w_{i1}; ...; u_{im}, w_{im}) = \sum_{j=1}^{m} f(u_{ij}) \times w_{ij} \tag{3-10}$$

where $f(u_{ij}) = \log_{10}(au_{ij} + 1)$ if $u_{ij} \geq 0$ and $f(u_{ij}) = -\log_{10}(-bu_{ij} + 1)$ if $u_{ij} < 0$ and $a < b$. Equation (3-10) replaces Equation (3-2) and (3-3) in the main equation (Equation (3-1)) such that:

$$U(X) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} f(u_{ij}) \times w_{ij} \right) \times \phi(t_i) \tag{3-11}$$

Equation (3-11) is more flexible by allowing an outcome to have multiple values or by allowing multiple outcomes for a given time. This flexibility can be used to model uncertainty in the value estimation for documents. First, let us assume that the information need of the user is uncertain, in which case multiple values can be assigned to a given document. For instance, a query '*jaguar*' may refer to a car model or an animal. Let us assume that the relevance value of a given document is 0.70 if it refers to a car and 0.30 if it refers to an animal. If the probability that the query refers to a car is 0.8 and to an animal is 0.2, the expected value of the document is $(0.8 \times 0.70) + (0.2 \times 0.30) = 0.62$. The probability of the query context could be estimated by analyzing the query logs of the user with additional information about its context. The estimation process is possible but expensive. Moreover, if the query is assumed to have a single context, the probability of the query context is 1.0 and the equation (3-11) is still applicable.

Now let us assume that the intertemporal choice model is interpreted such that there are multiple documents at a given time, which gives multiple values for each time *t*. For instance, there are multiple documents displayed at each time period in a browsing path. Therefore, each time *t* consists of multiple values generated by multiple documents. Equation (3-11) is applicable to compute the expected value for each time *t*. Without prior knowledge of the weight of the documents, an equal probabilistic weight can be

assumed for all documents in time *t*. In addition, a probabilistic weight of a document can be estimated based on its position (its rank) when it is displayed, such as the probability that a document at a given rank is relevant. The probability estimation can be done based on the distribution of the relevant documents in a given rank or by heuristic, such as by assuming that the probability of relevance for the documents decreases as the rank increases. The heuristic approach for the probabilistic weight estimation is investigated in Section 5.2.2.

3.3.4 Discount function

The discount function models the trade-off in receiving delayed outcomes. In the case of browsing, the delay exists due to the sequential assessment of the documents. In (Loewenstein and Prelec, 1993), the authors suggest that there are two motives affecting discounting, *impatience* and *preference for improvement* as described in Section 3.2.1.2. The impatience motive leads to a *positive time preference* and the preference for improvement motive leads to a *negative time preference*.

The main motivation behind any information seeking effort is the need to fulfill the information needs of the user. Usually, it is driven by the need for information to complete certain tasks. The suitability of the type of discounting chosen depends on the nature of the tasks.

3.3.4.1 Positive time preference

Positive time discounting is suitable when the user is impatient to complete his/her information seeking task. It is applicable if the task is constrained by time or cost. For instance, the task may require the user to find the relevant information as soon as possible, such as finding relevant information as a preparation for a meeting or for writing an urgent report. In this particular situation, there is *urgency* in completing the information seeking task. Therefore, the user is impatient, which leads to positive time discounting.

Moreover, sometimes the information seeking process is costly in term of time or money, such as searching from mobile devices. Every additional download of a document will incur additional cost to the user. Since every interaction is costly, a user may want to reach *reasonable* information as soon as possible. In this case, a trade-off between finding the relevant information and minimising the cost should be achieved.

It is normal for an individual to spend less time and less money to get what he/she wants. When an impatient user approaches an IR system to find information, he/she wants to find it as quickly as possible. Positive time preference supports this assumption.

3.3.4.2 Negative time preference

The second motive related to the discounting is the *preference for improvement.* In economic literature, there are a few cases where an individual favours the negative time preference such as the preference for an increasing salary, an increasing quality of life or an increasing health condition as time increases (Loewenstein and Prelec, 1991, Loewenstein and Prelec, 1993, van der Pol and Cairns, 2000). This means that an individual prefers to spend more or to sacrifice more now for a better future.

In Section 3.4.1, it is mentioned that the positive time preference is more appropriate for browsing due to the urgency and due to cost minimisation. In the case where there is no restriction in time to spend and the cost can be ignored, a user may want to do more exploration while browsing, such as a casual browsing of the Web to increase his/her knowledge about some topic. Without constraints, a user may want the system to lead him/her to a better document in term of relevance in the next iteration. This assumption is normally used by a relevance feedback system, where the user's feedback (the interaction) is used to improve the quality of the retrieved documents in the next iteration. Therefore, those who understand the mechanism of such a system will expect a better set of documents to be offered when they provide more feedback. This leads the user to expect that the value of delayed documents should increase as the time increases. For such users, a negative time preference may be more appropriate.

3.3.4.3 Shape of the discount function (exponential vs. hyperbolic)

The shape of a discount function used in the intertemporal choice model is also considered as one of the main properties discussed in the literature. The conventional assumption is that each time period has a constant discount rate which leads to an exponential discount function. It means that the value of a document is discounted at the same rate for the whole period. For instance, the value of document at time t_{i+1} is

half of the value at time t_i and the value of document at time t_{i+2} is half of the value of document at time t_{i+1} and so on.

However, in recent literature, researchers found that the assumption does not hold since it doesn't fit the real data collected from empirical studies (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Lazaro et al., 2002). Based on this data, it seems that a hyperbolic discount function is more applicable. This discount function assumes a declining discount rate as time increases.

In IR, there is no evidence to suggest which shape of discount function is more suitable for browsing. However, based on the findings in many empirical studies in economics, there is every indication that the hyperbolic discount function could be applicable to browsing as well. It is based on the assumption of the DU discussed in Section 2.6.1, where the discount rate of individuals is independent of the *consumption*, for instance, the discount rate for investment plan is the same as the discount rate for gambling. Nevertheless, this research question will be investigated in the experiment conducted in Section 5.4. In addition, the experimental setup to evaluate the intertemporal choice model for browsing in this thesis is designed based on assuming that the *positive time preference* is used. Therefore, this thesis will only focus on positive time preference for browsing.

3.4 Discussion

This research contributes to some development in the IR research especially in the context of interaction. First, this research explores the potential of using the intertemporal choice model to model browsing behaviour of users in an attempt to provide a solution to the browsing problem, which is to find the optimal browsing strategy for the user. Second, the concept of discounting for future documents in the intertemporal choice model is compared to the existing discount function applied to documents. Finally, the value function incorporates two psychological properties of individual decision making in the model, which potentially allows the system to better model the interaction of the user.

3.4.1 Modelling browsing behaviour

The application of decision theory in IR research is not a new development in this field. The expected utility theory (EU) (von Neumann and Morgenstern, 1944) has been studied for indexing (Cooper and Maron, 1978), retrieval (Salton and Wu, 1980) and evaluation (Bollmann and Raghavan, 1988) of IR systems. Through decision theory, many aspects of an IR system can be modelled, such as how to choose documents for retrieval, how to choose good index terms for a document, how does a user interacts with the system *et cetera.* Decision theory provides a good foundation for modelling these decision problems in IR.

In this thesis, the intertemporal choice model, a type of decision theory, is adopted to model the browsing behaviour of the user. It is based on the assumption that the browsing behaviour is an instance of an intertemporal choice problem. The sequence of documents in browsing is viewed as the set of delayed documents for the user. The intertemporal choice model has been investigated extensively in economic and social choice studies (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Frederick et al., 2002, Lazaro et al., 2002). Based on empirical studies, the usefulness of the model has been proved. It is the aim of this research to investigate the effectiveness of the intertemporal choice model in modelling the browsing behaviour of the user in general, even though the evaluation focuses on the application of the model to the problem of searching by using mobile devices and to the problem of subtopic retrieval. Moreover, by adopting a descriptive model of decision theory, where the model is developed based on empirical data, the model should be able to capture the actual behaviour of the user when interacting with the IR system, thus it should be able to find the optimal strategy for browsing.

In addition, applying the intertemporal choice model to IR problems also serves as a contribution to the intertemporal choice model itself. Most research on the model is concentrated on economic and social choice problems where most of the outcomes are monetary. In this case, the research attempts to observe the effect of the intertemporal model in another domain where information is the currency.

Furthermore, in Section 3.3.3.4 the intertemporal choice model is extended to model uncertainty on the value estimation of the outcomes. This modification of the model allows an outcome to have a set of values with probability of uncertainty associated to

the values. This feature is quite useful in the context of IR, since the value of information is not always quantifiable.

Finally, the intertemporal choice browsing model could be a better way to model user browsing behaviour. First, it eliminates the need to collect historical browsing data for modelling the browsing behaviour. One way to model user browsing behaviour is based on the browsing history. However, the IR system rarely records the browsing history of users. In addition, the information from Web access logs is not sufficient for modelling browsing behaviour since no task or relevance judgement is recorded for the browsing data. On the other hand, the proposed intertemporal choice browsing model could model user browsing behaviour by treating browsing as an intertemporal choice problem. Therefore, the problem of modelling browsing behaviour is reduced to the problem of intertemporal choice in IR. Second, it has been suggested that there is a threshold value used by a user to decide when choosing the next document to browse or to read (Huberman et al., 1998). A document will be chosen next if the value of reading the document is higher than this threshold value. Therefore, in order to model browsing behaviour of users, such a threshold value should exist and be known. However, modelling the threshold value is challenging. In the case of the intertemporal choice browsing model, the psychological aspect of decision in browsing is captured by a suitable discount function, which has been extensively studied elsewhere (Cairns and van der Pol, 2000, Frederick et al., 2002, Lazaro et al., 2002). Hence, the need to model the threshold value in browsing is eliminated.

3.4.2 Discounting of the future documents

One of the properties of the intertemporal choice model is the discounting of future (or delayed) documents. In IR research, the use of discount function for documents has been investigated in the context of relevance feedback models. The ostensive model (Campbell and van Rijsbergen, 1996) is one of the relevance feedback models that uses such a discount function. In the model, the weight of a document is discounted based on its *age*. The age of a document in a browsing path is defined as the number of documents separating the document and the current document. For instance, if a browsing path consists of four documents, the age of the first document in the browsing path is three, calculated with respect to the current (last) document. There are three profiles proposed; the weight of a document decreases, is constant or increases as the age increases. Campbell claimed that the decreasing with age profile is the best

(Campbell, 2000). Hirashima et al. proposed a similar discount function for browsing in hypertext (Hirashima et al., 1998). They made the same claim that the decreasing function with age is the best.

The discount function in the intertemporal choice model is similar to the one proposed in the ostensive model. In the context of positive time preference, the discount function is similar to the decreasing with age profile and vice versa. Imagine that a user is at a current time t_0. In the case of the ostensive model, the discounting is applicable to previous documents, while for the intertemporal choice model, the discounting is applied to future documents. However, both share the same concept where the discounting is determined by the distance (in term of time) between the documents with respect to t_0, where the intertemporal choice model assumes a forward time direction and the ostensive model assumes a backward time direction.

3.4.3 Value function

The value function of the intertemporal choice model brings two new developments into IR research, the transformation function for the document's value and the sensitivity toward different assumptions used for estimating the document's value. With these new developments, this research explores the possibility of modelling user behaviour whilst interacting with an IR system.

In Section 3.2.1.1, the concept of wealth and gains/losses as part of the value function in the intertemporal choice model has been described and, in Section 3.3.3, the concepts of collective and differential value have been proposed as alternatives to the wealth and gains/losses. Assuming that people will be sensitive to the collective and differential value as they are to wealth and gains/losses, it is assumed that an IR system will be more effective from the user's perspective when adopting these concepts for modelling the value of a document.

Both concepts assume that the value of a document depends on the other documents, which is also known as dependence relevance. The concept of dependent relevance is seldom used in IR models but it has been investigated since the early 70's. Goffman's indirect retrieval method assumed the dependent relevance concept (Goffman, 1969). The method assumes that the relevance of a document depends on other documents. The next document to be retrieved should be the most relevant with respect to the

immediate previous document retrieved regardless of the query. An example of its implementation can be found in (Croft and van Rijsbergen, 1976).

In Section 3.3.3.1, it is assumed that the collective value of a set of documents could be measured as a group entity, instead of for individual documents. This is similar to cluster based retrieval (Croft and van Rijsbergen, 1976), where each cluster (a group of documents) is represented by a cluster representative (such as the centroid). During retrieval, each cluster is assigned a score based on the similarity of the cluster representative and the query. The cluster representative can be seen as an aggregate document. In another approach, the value of a document is computed based on the similarity of an aggregate query to the document. The refinement of the original query based on the documents already viewed by the user in the ostensive model (Campbell and van Rijsbergen, 1996) can be regarded as a type of aggregate query.

In Section 3.3.3.2, it has been suggested that novelty models can be used to estimate the differential value of documents. The novelty score of a document reflects the amount of new information contained in the document. A number of good techniques have been proposed to model the novelty score (Zhang et al., 2002, Allan et al., 2003, Gabrilovich et al., 2004). It is assumed that existing models could be adopted to compute the differential value of a document. The novelty score alone may not be useful if it does not reflect the information need of the user. Therefore, researchers have investigated the techniques to combine the novelty score and the relevance score of a document. For instance, the Maximal Marginal Relevance (MMR) ranking method proposed a linear combination of the two scores for ranking the retrieved documents (Carbonell and Goldstein, 1998). Alternatively, a cost based combination of the two scores was proposed by Zhai *et al.* for their Risk Minimization Framework (Zhai et al., 2003). In this thesis, the use of different novelty models and the MMR ranking technique to estimate the dependence value for a document in the context the intertemporal choice model is investigated.

3.5 Chapter Summary

In Section 3.2, an intertemporal choice model that is often used to model decision making behaviour with respect to choices of outcomes that are spread over times is described. Since the model is successful in many economic applications, the model will be used to model decision behaviour in browsing. In Section 3.3, the issues and

solutions for adopting the intertemporal choice model to model interaction in IR were discussed. In Section 3.4, some contributions of the proposed model with respect to the existing research work were presented. In the next three chapters, the application and the evaluation of the intertemporal choice model will be discussed. In Chapter 4, an implicit relevance feedback (RF) system is evaluated as an IR system for mobile devices. Then, a recommendation system based on the intertemporal choice model to improve the performance of the implicit RF system is evaluated in Chapter 5. After that, the use of the intertemporal choice to provide an optimal ranking of retrieved documents in the context of the subtopic relevance retrieval application is evaluated.

4 An Implicit Relevance Feedback System for Mobile Devices

4.1 Introduction

This chapter describes and investigates an interactive IR (IIR) system for searching for information by using mobile devices. The architecture of the system is based on the system proposed in (Vinay et al., 2005). Searching and browsing are part of the information seeking strategies used and supported by the system. The intertemporal choice browsing model discussed in Chapter 3 will be incorporated as the browsing recommendation model for the system to recommend an optimal browsing strategy to the user, which will be discussed in the next chapter.

The aim of this chapter is to investigate the performance of the IIR system proposed in (Vinay et al., 2005) in helping the user to search for information on mobile devices. The system is chosen mainly because of its simple browsing approach and also its suitability to incorporate the intertemporal choice browsing model investigated in this thesis as well as its suitability to evaluate the model. Such a system focuses on maximising the retrieval performance as well as minimising the cost of the retrieval. In this chapter, the evaluation of the system includes finding the upper and lower bound of the system's performance before the incorporation of the intertemporal choice browsing model into the system. Different configurations for the IIR system will also be investigated in this chapter and the best configuration is sought. The configuration includes the *display strategy*, that is, the strategy to select the documents to be displayed

to the user, and the *expansion model*, that is, the model used to retrieve the relevant documents based on user's feedback, used in the system.

The advancement of mobile technology enables us to access information anywhere and anytime. With internet-ready mobile devices, such as mobile phones and PDAs, searching for information becomes much easier. The user may want to use a search engine while on the move, such as on the bus or on the train. Thus, the search engine might become one of useful applications on mobile devices in the near future.

However, there are two limitations of mobile devices that will affect the effectiveness of the system, namely their small screen size and their limited and expensive interaction capability. The small screen size makes it difficult to display a normal Web page properly. Extensive scrolling is usually required in order to read the pages and it is problematic for the user. Some researchers have studied the use of summaries to help the user reading the Web pages (Buyukkokten et al., 2002, Sweeney and Crestani, 2006). The number of search results that can fit on the relatively small screen is limited. Therefore, scrolling is inevitable. Due to this fact, Vinay *et al.* chose only *four* documents from the retrieved set to be displayed on the screen each time (Vinay et al., 2005).

The limited interaction capability of the mobile devices makes it difficult for the user to use it for searching. For instance, typing a query on the keypad of a mobile device is problematic. The system should minimize the need for extensive user interaction on mobile devices. Therefore, it is suggested in (Vinay et al., 2005) that the system should be able to interpret the user's information needs based on his/her interaction. In particular, the proposed system makes the assumption that the documents chosen by the user are all relevant and it uses a relevance feedback (RF) model to infer his/her information need. As a result, the system may display a better set of documents to the user in the next iteration.

The remainder of the chapter is structured as follows. In Section 4.2, the architecture of the IIR system will be described in details. The evaluation strategy for the system will be discussed in Section 4.3. The results of the experiments will be presented in Section 4.4. The discussion of the experimental results will be given in Section 4.5. Finally, the summary of this chapter will be given in Section 4.6.

4.2 An Interactive IR System for Mobile Devices

In (Vinay et al., 2005), the authors proposed an interactive IR system for small display devices. The system combines a relevance feedback (RF) module and a display strategy module to provide an interactive browsing tool for searching documents. The display strategy module is used to select a certain number of documents from the retrieved set to be displayed to the user. When a user chooses to read one of the displayed documents, the system treats this as positive feedback from the user. The RF module learns from the feedback and retrieves another set of documents for the next iteration. The process continues until the user stops after he/she finds the intended information or after he/she gives up.

Based on this algorithm, the problem of searching for documents on a mobile device can be overcome by displaying fewer documents each time to fit on the small screen of the devices and by learning from the user's interaction (e.g. selecting a document) to guide the search. Therefore, the system proposed a reasonable solution to the problem of searching for documents on mobile devices.

Figure 4-1 shows the conceptual model of the interactive IR system. There are six components of the system, which are the user, the query, the IR system, the display strategy, the interaction and the expansion model. It comprises two cycles, the query cycle indicated by the black arrows and the browsing cycle indicated by the red arrows. The dashed line between the IR system and the expansion model indicates that there is a hidden communication between the two objects in the browsing cycle.

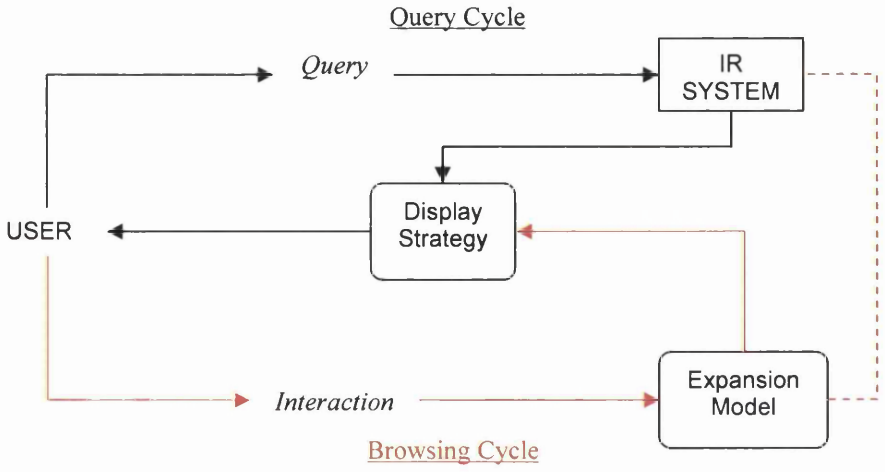


Figure 4-1: The interaction model in the interactive IR system

Usually, the user starts by submitting a query that best represents his/her information need to the IR system. The IR system finds and retrieves all documents that are judged relevant to the query. The display strategy will determine how the retrieved documents are displayed to the user. A common strategy is to display the top-*k* of the retrieved documents to the user where *k* is usually *10* or *20*. Once the documents are displayed, the query cycle is completed.

Upon receiving the retrieved documents, the user assesses them to find the required information. The user may decide to reformulate his/her query and resubmit it to the IR system and another query cycle will occur. Or, the user may choose to browse by selecting one of the displayed documents. In this system, the interaction is interpreted as feedback from the user. The expansion model will infer the needs of the user from his/her interaction and will suggest another set of documents to be retrieved. For instance, in the case of the RF model, the user query is refined based on the user's feedback and the refined query is submitted back to the IR system to retrieve another set of documents. The retrieved documents are filtered by the display strategy before they are presented to the user and the browsing cycle is completed. Usually, the user will continue to interact with the system (by querying or browsing) until he/she is satisfied with the information he/she received or he/she gives up.

Note that the expansion model and the display strategy are additional components included into the conventional IR system. The expansion model acts as an adaptive agent that interprets user's interactions and suggests an action from the IR system which is usually to retrieve another set of documents to satisfy the user information need. In this case, the set of retrieved documents depends on the displayed document chosen by the user. The set of retrieved documents are generated based on the reformulated query submitted to the IR system.

The display strategy has been part of the conventional IR system and it has been assumed that displaying the top-*k* retrieved documents is always the best strategy, following the Probability Ranking Principle (PRP) (Robertson, 1977, Gordon and Lenk, 1991). However, due to the fact that the outcome of the expansion model depends on the documents chosen by the user, displaying the right set of documents to the user becomes an important strategy. In the case of *active feedback* (Shen and Zhai, 2005),

those documents that the system is most unsure of being relevant or not are displayed to the user. Obtaining feedback on these uncertain documents will make the process of discriminating the relevant evidence from the non-relevant evidence a lot faster. For this system, the display strategy will either make the browsing converge to a specific topic or diverge to a broader topic. A browsing session will be a success if it converges to the intended topic or will be a failure if it does not. Therefore, the display strategy plays an important role in the performance of the system.

In the following sections, two different display strategies and an expansion model that can be used in the system are described. In the case of the display strategy, the strategies proposed for the active feedback system in (Shen and Zhai, 2005) are adopted, which are the top-k documents and the gapped top-k documents. Another strategy, which is to use the centroid of the k-clustered documents for the display strategy is not implemented, due to the fact that clustering the set of retrieved documents before displaying them to the user is not an efficient solution.

4.2.1 Display Strategy

The research problem in terms of the display strategy is to decide which documents should be displayed to the user in order to improve the effectiveness of the system. The effectiveness of the implicit RF model in inferring the information need of the user depends on the documents selected by him/her. In addition, the user can only choose a document from the set of documents displayed to him/her. Therefore, the set of displayed documents can affect the effectiveness of the system. The problem of finding the right strategy to display or to return the documents to the user is similar to the problem of active feedback (Shen and Zhai, 2005).

Displaying the top-k documents is a common display strategy. The strategy assumes that most of the relevant documents will be highly ranked by the IR system. Therefore, it will increase the likelihood of the user finding relevant documents if the top-k documents are displayed to the user. The strategy of displaying the top-k documents is depicted in Figure 4-2(a).

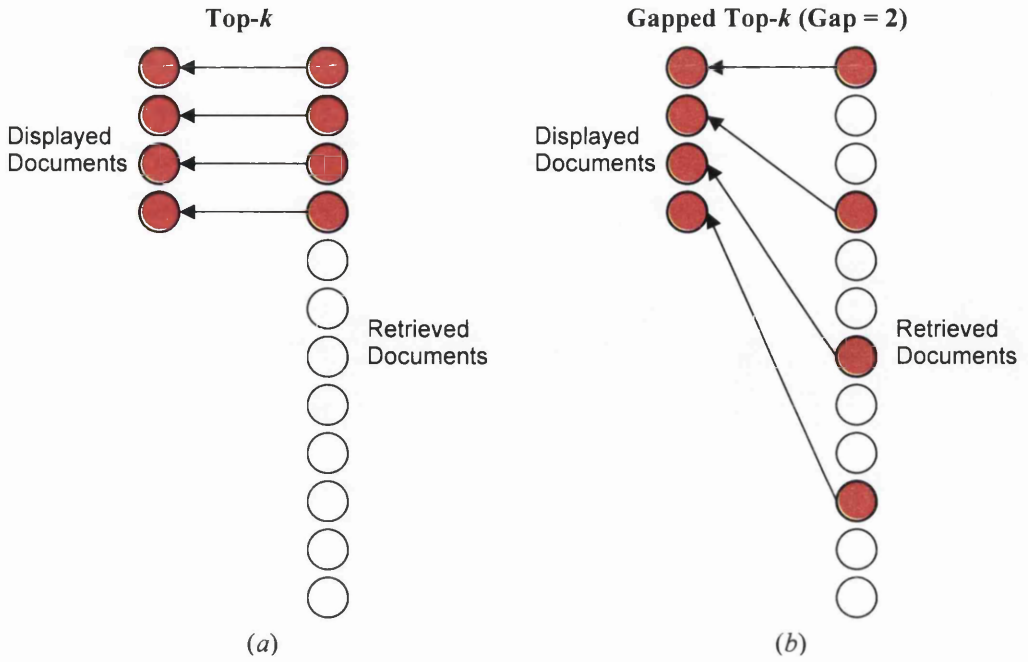


Figure 4-2: The top-*k* and the gapped top-*k* display strategy

Due to the fact that the documents ranks are computed based on a common query, there is a high possibility that the top documents are similar in content. Assuming that the user will choose one of the displayed documents and the system will treat it as relevant, there is a high chance that the next set of retrieved documents will be of the same topic, because the refined query will most likely be similar to the content of the chosen document. If the user's information need is concerning only this single topic, then this is a good strategy. But, if it is not about the topic that the user intended, or if the user's information need covers multiple topics, the browsing session will be a failure. Therefore, displaying the top-*k* documents may not always be an effective strategy.

In order to prevent such situations, the system should display a set of documents that covers a broader topic or content, such that it will give the user a better coverage of the topic. A user will have an option to look for a very specific sub-topic by choosing one of the displayed documents that is closely related to the sub-topic. It may give an opportunity for the user to find documents that are not about the topic covered by the top-*k* documents. Two of the techniques to solve the problem proposed in (Shen and Zhai, 2005) are considered, which are the gapped top-*k* documents and the centroid of the *k*-clustered documents.

The gapped top-*k* strategy is shown in Figure 4-2(b). The documents to be displayed are selected from the top ranked documents. However, instead of choosing the top documents, the strategy is to select the documents from the top with a given gap. For instance, if the gap is *two*, the documents to be displayed are those at rank 1, 4, 7, and 10, as depicted in Figure 4-2(b). This is a simple and efficient technique to generate a set of documents with broader topic coverage.

The technique of choosing the displayed documents based on the centroid of the *k*-clustered documents is illustrated in Figure 4-3. A set of the top retrieved documents are clustered into *k* closely associated documents. The contents of the documents in a cluster should be more similar to the other documents in the same cluster than to the documents in other clusters (Van Rijsbergen, 1979). The centroid of the cluster, illustrated as the red circles, represents the content of the documents in the cluster and, in this case, the centroid is one of the documents in the cluster. Assuming that each cluster represents a topic, there are *k* topics contained in the top retrieved documents. The user will be given a broader choice of topics by displaying the centroids of the clusters to him/her.

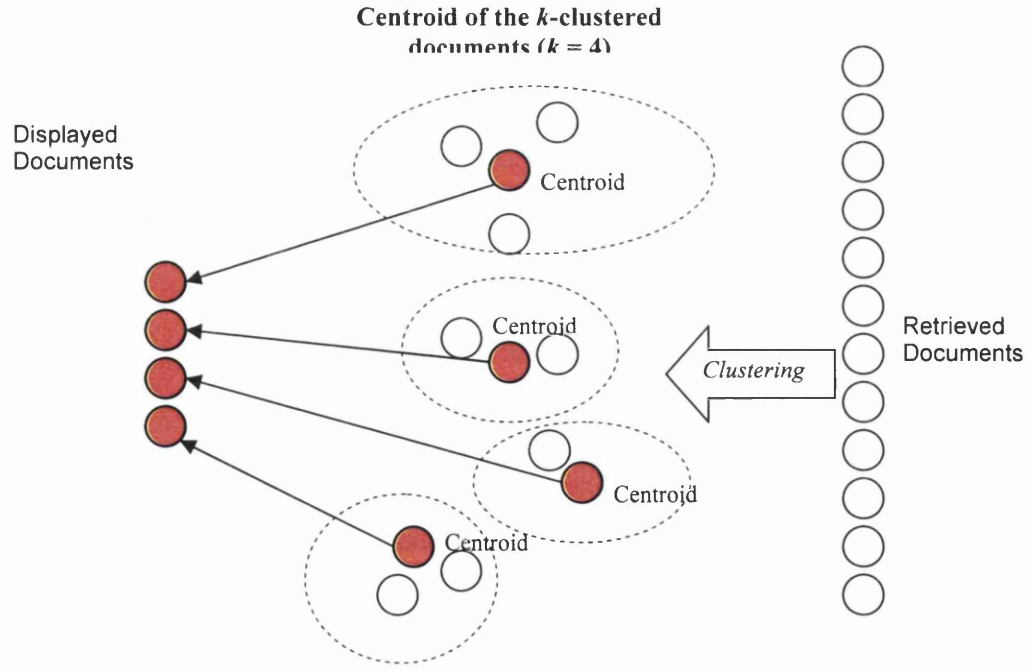


Figure 4-3: The centroid of the *k*-clustered documents

In this chapter, only the gapped top-*k* documents technique is adopted for the display strategy. For the second strategy, the retrieved documents need to be clustered into several clusters of documents before the selection for the documents to be displayed can take place. The clustering of the documents is computationally expensive and it makes the system inefficient. Therefore, it is decided that the second technique is not suitable for this application.

4.2.2 Expansion Model

The role of the expansion model is to choose a set of documents to be retrieved for the next iteration based on learning from the interaction of the user. The new set of retrieved documents will be the input to the display strategy module to choose which documents to be displayed to the user, as shown in Figure 4-4.

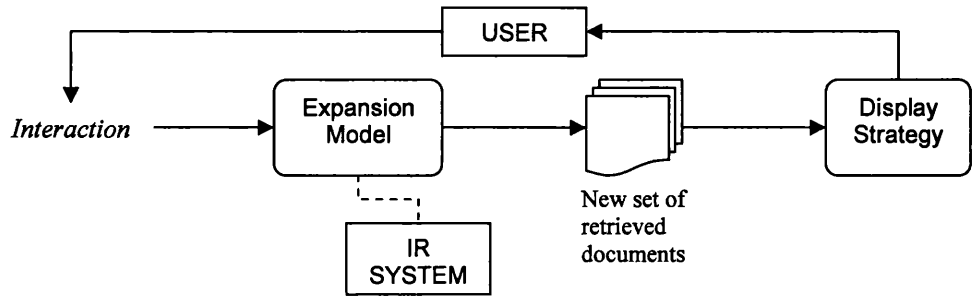


Figure 4-4: The mechanism of the expansion model

A simple strategy is to use the remaining retrieved documents that are not displayed as the candidates for the next iteration. In this case, the document chosen by the user will not affect the documents selected for the next iteration. It is similar to the effect of viewing the retrieved documents page-by-page. Choosing any document is similar to clicking a next-page button.

The next strategy is to use an implicit relevance feedback (RF) method to learn from the feedback of the user and to retrieve another set of documents based on the refined query. Such a strategy is depicted in Figure 4-5. In this case, the documents chosen by the user in the previous iteration are treated as relevant to the user's information needs (Campbell and van Rijsbergen, 1996, White et al., 2004, White et al., 2005). Therefore, a new and refined query is generated based on the feedback.

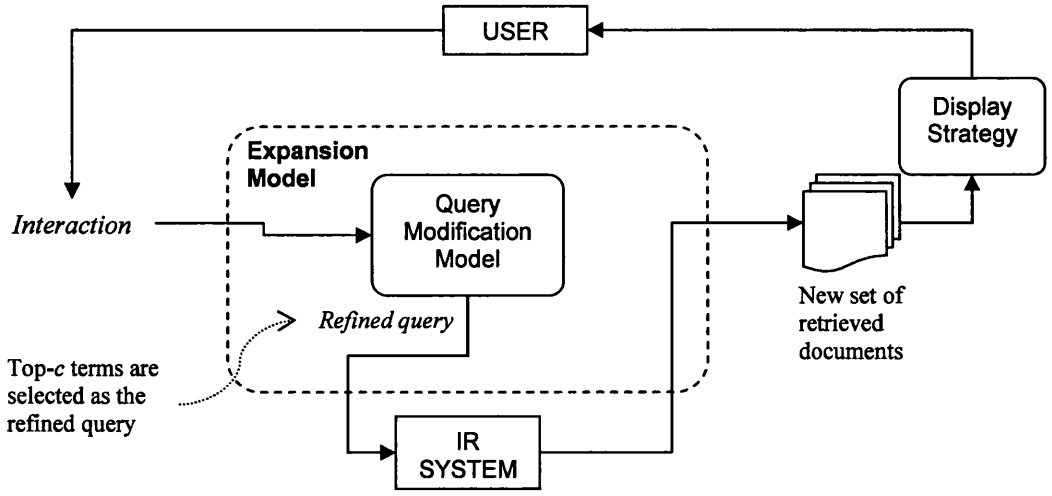


Figure 4-5: The expansion model

Usually, the feedback information is used to assign weights to all index terms in the collections and to rank the index terms according to their weights. The top-*l* terms are chosen as the new refined query. The computation of the terms' weights is given by the RF model. The use of the *ostensive model* (Campbell and van Rijsbergen, 1996, Campbell, 2000, Urban et al., 2003) as the implicit RF model will be investigated in this chapter.

The main advantage of the implicit RF model is that the refined query is generated based on the feedback from the user. The refined query should be a better representation of the user's information need as his/her need may develop while browsing (Campbell and van Rijsbergen, 1996). Moreover, the documents already read by the user while browsing define the context of his/her information need (Hirashima et al., 1998). Therefore, a better set of documents will be retrieved by using the refined query.

The ostensive model is a path-based implicit RF model. The model assumes that the information need of the user are developed as he/she sees more documents during an interactive browsing session (Campbell and van Rijsbergen, 1996, Campbell, 2000, Urban et al., 2003). The refined or developed information need of the user can be simplified as a set of weighted index terms. The weight of a given index term is estimated based on its distribution in the documents in the path.

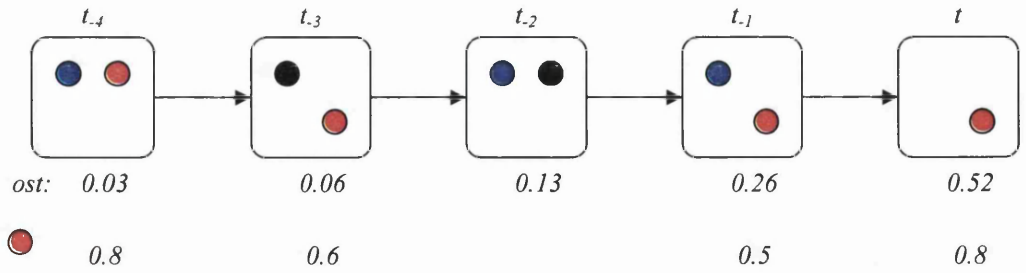


Figure 4-6: An example of a browsing path

Figure 4-6 shows an example of a browsing path of five documents. The documents are represented as by the rounded rectangles and the arrows represent the sequence of the browsing path. The coloured circles represent the index terms and the index terms are assumed to be binary. If a coloured circle exists in a given document, then the index term appears in the document. In the figure, a decreasing *ostensive profile* is assumed where the weight of a term in a document decreases as the age of the document increases. The right-most document is the most recent document and the age of a document increases as its distance from the most recent document increases (left-wise direction) as shown in the figure. In the example, the ostensive weight for each document is shown below the document. The value for the '*red*' index term is also shown below the respective document.

$$weight(red) = (0.03 \times 0.8) + (0.06 \times 0.6) + (0.13 \times 0.0) + (0.26 \times 0.5) + (0.52 \times 0.8) = 0.61$$

Figure 4-7: The computation of the index term's weight by using the Ostensive Model

The computation of the weight of the index term is shown in Figure 4-7. In this example, the final weight of the index term is equal to 0.61. Once the weight of all index terms are computed, the top-l index terms are chosen as the refined weighted query. In this chapter, the effectiveness of three ostensive profiles, the *decreasing*, the *constant* and the *increasing* profiles (Campbell, 2000) for the ostensive model based implicit RF system will be investigated.

4.3 Experimental Methodology

Evaluating an interactive information retrieval (IIR) system is an expensive task. The experiment usually involves real users using a real system with specific information seeking tasks. It is time consuming and it requires a lot of effort from the researchers and from the subjects; the users recruited to participate in the experiment, before, during

and after the experiment. Therefore, any experiment involving users should be done only if necessary and it should be well-planned.

A non-user evaluation method for an IIR system can be achieved through simulation. White *et al.* show that some features of an IIR system can be evaluated by using simulation-based studies (White et al., 2004, White et al., 2005). In the studies, they simulate user interactions by generating a set of possible browsing paths that the users may use and evaluate the performance of the system based on these paths. Similarly, the evaluation can be conducted based on the behaviour of the system. Vinay *et al.* demonstrate that a set of system's responses based on all possible interaction decision of the users can be used to evaluate the overall effectiveness of an IIR system (Vinay et al., 2005). The evaluation in this chapter is based on the simulation-based studies conducted in (Vinay et al., 2005).

The main focus of this simulation-based evaluation is to investigate the effectiveness of an IIR system based on the quality of the retrieved documents and not based on the quality of the system interface. The interface could only be evaluated by real users. However, a user study would not be able to separate interface issues from retrieval performance issues easily. Therefore, the experimental IIR system is designed and implemented with such aim in mind.

The aim is of this evaluation is to discover the performance of the IIR system in an information seeking task in general, such as how many relevant documents it can retrieve and what the chances are that the user will find the information required to make his/her information seeking session successful. It is different than the ad-hoc retrieval evaluation in the sense that the performance is measured based on the entire information seeking (browsing) session of the user, which includes multiple retrieval iterations, rather than on the quality of the set of retrieved documents in single retrieval iteration.

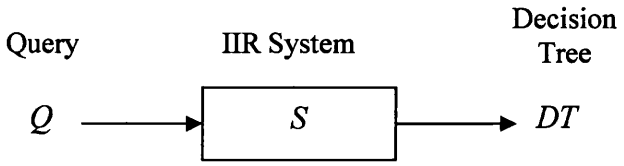


Figure 4-8: The evaluation model for an IIR system (an implicit RF system)

Figure 4-8 shows the evaluation model for an IIR system. The model consists of a *Query* (Q), an *IIR system* (S) and a *Decision Tree* (DT). A query in this model is the initial query statement formulated and submitted by the user to an IIR system. It is not the *refined query*[2] that may have been generated and resubmitted to the IIR system within the interactive session such as in the case of the relevance feedback (RF) system.

Once a query is submitted, the IIR system responds by displaying a set of documents to the user. A given user is allowed to interact with the system by choosing one of the displayed documents. The action of selecting a document is treated as feedback and the system will respond by displaying another set of documents that it judges to be relevant. As a result, the system will iteratively select a set of documents for a user to assess. The continuous process of document assessment by a user is considered browsing. The browsing session starts with the query submitted by a given user and an iterative assessment of documents until the user is satisfied.

A decision tree consists of all possible documents displayed by an IIR system responding to all possible interactions of the user for a given query. Figure 4-9 shows an example of the decision tree that is generated by fixing the size for the set of displayed documents to *four* and the number of iterations to *five*. Based on the decision tree, we can observe the behaviour of the IIR system responding to a given query and also to any possible actions (selecting a document) by the user.

In the context of the evaluation, for each query submitted to the IIR system, there will be a decision tree that shows all possible outcomes of the browsing session. The performance of an IIR system is evaluated based on the decision tree generated by the system. This is different to the ad-hoc task in which the system is evaluated based on the set of documents retrieved by the system in response to initial query. Indeed, the decision tree simulates the potential interaction behaviour of the user and the IIR system.

[2] The new query generated after the refinement process

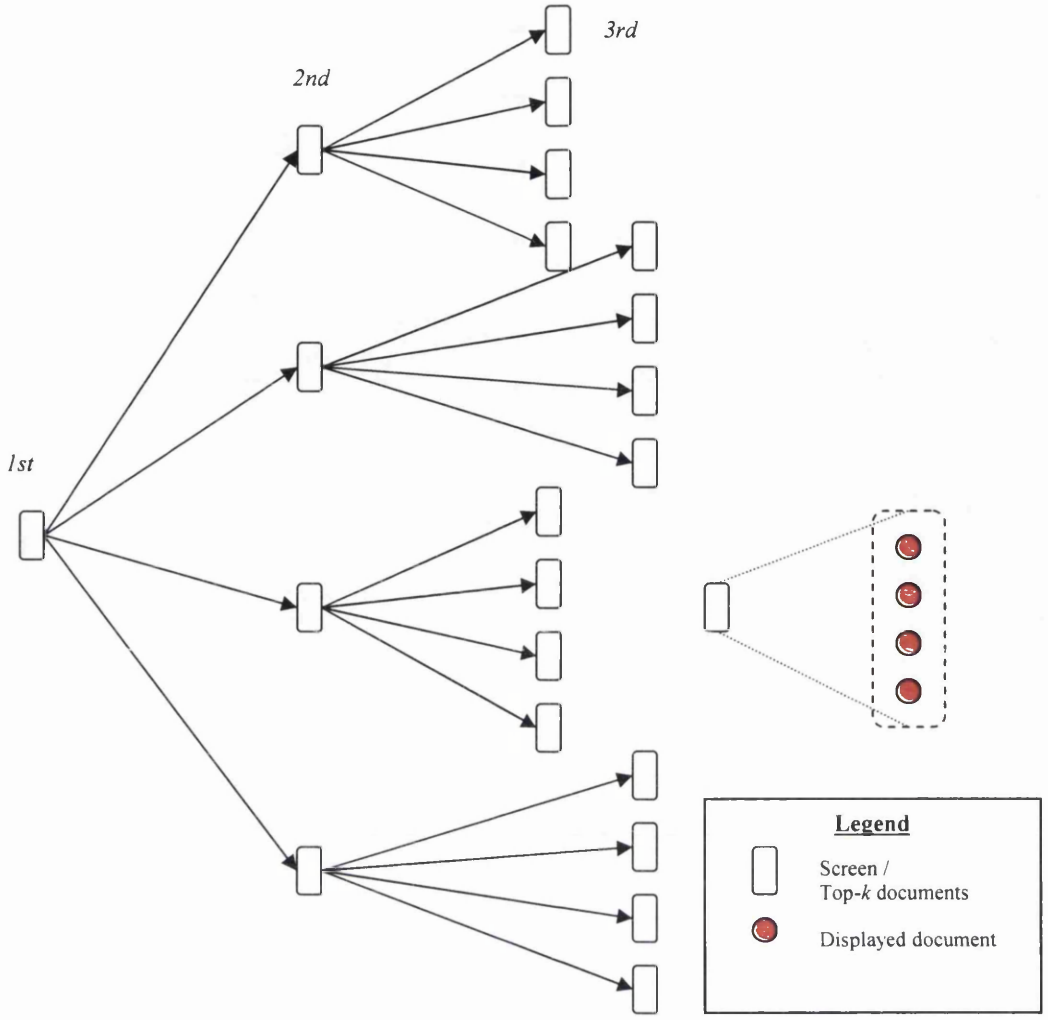


Figure 4-9: An example of a decision tree produced by the implicit RF model

4.3.1 Evaluation measures

In (Vinay et al., 2005), the task in the evaluation is to find a specific document by using a set of randomly chosen words from that document as the query. In the evaluation, the rank of the document in two retrieval strategies, the conventional top ranking system and the alternative iterative ranking system based on relevance feedback, is compared. The aim is to show that the iterative ranking system should push up the rank of a given document, thus making it easier for the user to find it.

However, in many cases the user does not seek a known document. He/she may not even know which document could satisfy his/her information need. Sometimes, it is hard to formulate a good query statement to represent his/her information need. In addition, his/her information need can only be satisfied by more than just one document. Therefore, evaluating the system based on the rank comparison for many relevant documents may not be suitable for this evaluation task.

The effectiveness of an IIR system cannot be measured based only on the documents retrieved in the first iteration. It should measure the overall performance of the system in responding to the interaction with the user in a given information seeking session. In this case, *precision* and *recall* measures may not be entirely accurate as a performance measure for an IIR system. In this thesis, the performance of the IIR system is evaluated based on the quality of the set of possible browsing paths generated by the system for each query. The quality of each browsing path is measured based on the *Mean Average Precision* (MAP) score or the *expected search length* (ESL) score.

4.3.1.1 Mean Average Precision (MAP)

An example of a possible browsing path extracted from a decision tree is depicted in Figure 4-10. In this case, the length of the browsing path is *five* and the number of documents displayed to the user is *four*. In Figure 4-10, the red circles represent the relevant documents while the white circles represent the non-relevant documents.

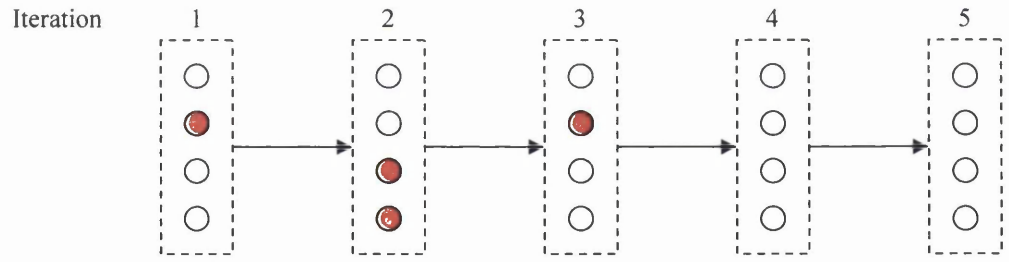


Figure 4-10: An example of a browsing path extracted from a decision tree

The mean average precision (MAP) score is usually computed for a ranking of documents and not for a browsing path. Therefore, a browsing path needs to be transformed to a ranking of documents before a MAP score can be measured. The documents in a browsing path are ranked based on their sequence of appearance; ranking the first set of displayed documents followed by the next set of documents. As such, the browsing path in Figure 4-10 will be transformed to a ranking of documents as depicted in Figure 4-11.

Figure 4-11 shows an example calculation of the MAP score for the browsing path. At each relevant document found in the ranking, a precision score is computed; the ratio of the number of the relevant documents found so far to the current rank number (or the number of documents assessed so far). The MAP score is the average precision scores

computed for each relevant document in the ranking. If there are *four* relevant documents in the ranking, the MAP score is the average of the *four* precision scores as depicted in Figure 4-11.

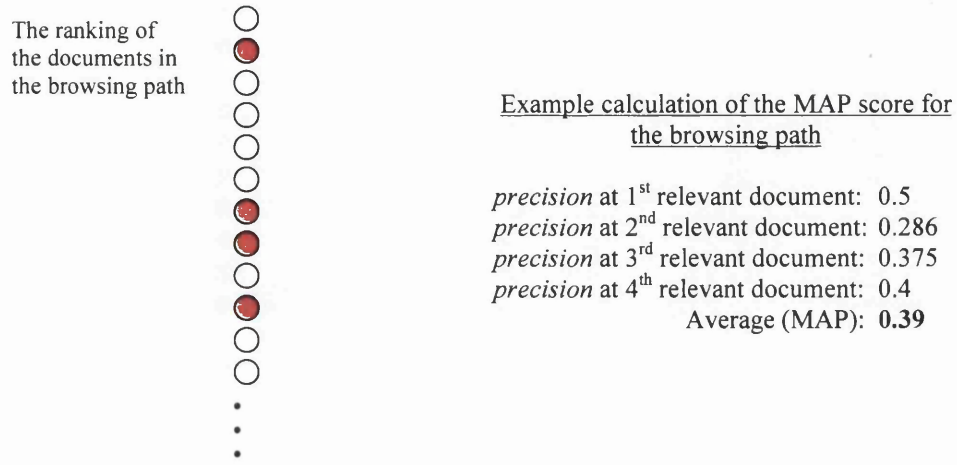


Figure 4-11: Calculation of the MAP score for the browsing path

The MAP score, often used in the retrieval evaluation, measures the distribution of the relevant documents in a ranking of documents. For instance, a higher MAP score can be expected if more relevant documents are ranked at the top and *vice versa*. In the case of browsing, more relevant documents at the top will indicate that less non-relevant documents need to be assessed. Therefore, the browsing path with a higher MAP score is a better browsing path for the user.

4.3.1.2 The expected search length (ESL)

The cost of browsing in an interactive IR system can be defined in many ways. One of the costs often discussed is the cost of downloading a document onto a mobile device such as a PDA or mobile phone. Assuming that the cost for downloading a document is equal regardless of its size, the downloading of a non-relevant document is unnecessary and a costly overhead. Therefore, an interactive IR system should minimize the number of non-relevant documents downloaded while the user is searching for relevant documents.

The *Expected Search Length* (ESL) was proposed to measure the effort or the cost of browsing through the set of retrieved documents (Cooper, 1968, Bollmann and Raghavan, 1988). The cost of browsing or the effort of the user is measured based on the expected number of non-relevant documents read before a certain number of relevant documents is found. According to (Cooper, 1968), the effectiveness of a

retrieval system should be measured based on how much the effort of the user can be reduced by using the system.

In (Cooper, 1968), the purpose of constructing a weak ordering of documents is to put constraints on the search such that the user needs to assess the documents in batches. Each batch of documents corresponds to a rank level. Each rank can consist of one or more documents and the user randomly assesses the documents in each rank. The ESL measures the expected number of non-relevant document to be found before a certain number of relevant documents are reached (Cooper, 1968).

A browsing path, such as the one in Figure 4-12, consists of five batches of documents. In this case, the search is constrained by displaying batches of documents iteratively to the user. Similarly, the documents in each batch are assumed to be assessed randomly. The batches of documents displayed by the system are considered a weak ordering of the documents. Therefore, an ESL can be calculated for each browsing path.

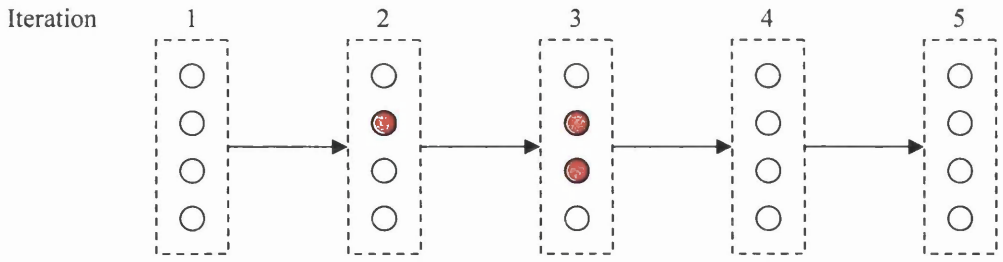


Figure 4-12: A typical browsing path (session) of a user

In this evaluation, the expected search length (ESL) of finding the *first* relevant document is used as another evaluation measure of the effectiveness of a browsing path (session). It is a measure of the expected number of non-relevant documents needs to be assessed before the first (*one*) relevant document is found in the browsing path. As such, the expected cost of browsing in a browsing path (session) is represented by the expected number of the non-relevant document assessed before the first relevant document is found.

Based on the example in Figure 4-12, there are four non-relevant documents in the first iteration. The first relevant document can be found in the second iteration, where there are three non-relevant documents in the iteration. Since first relevant document is only

found at the second iteration, the minimum number of non-relevant documents to be assessed is 4. Moreover, since there are other 3 non-relevant documents in the second iteration, the maximum number of non-relevant documents to be assessed is 7. Let *nr* be the number of non-relevant documents to be assessed, the probability of finding less than 4 non-relevant documents or finding more than 7 non-relevant documents is *zero*, such that $p(nr < 4) = 0.0$ and $p(nr > 7) = 0.0$ as depicted in Figure 4-13.

$$p(nr < 4) = 0.0$$
$$p(nr = 4) = 0.25$$
$$p(nr = 5) = 0.25$$
$$p(nr = 6) = 0.25$$
$$p(nr = 7) = 0.25$$
$$p(nr > 7) = 0.0$$
$$ESL_1 = (0.0) + (0.25 \times 4) + (0.25 \times 5) + (0.25 \times 6) + (0.25 \times 7) + (0.0) = 5.5$$

Figure 4-13: An example of ESL calculation for the browsing path when $i = 1$

Since the documents in the second iteration are randomly assessed, there are *four* possible numbers of non-relevant documents to be assessed, which are 4, 5, 6 and 7. Therefore, the probabilities of the non-relevant documents are estimated accordingly. If the first document chosen in the second iteration is the relevant document, the total number of non-relevant document already assessed is 4 (the documents in the first iteration) and the probability of randomly choosing the relevant document in the second iteration is 0.25, such that $p(nr = 4) = 0.25$. Similarly, if one non-relevant document is chosen first before the relevant document, then the total number of non-relevant documents is 5 (including the documents in the first iteration) and the probability of occurrence for such scenario is 0.25, such that $p(nr = 4) = 0.25$. The probabilities of the other scenarios are estimated based on the same principle and Figure 4-13 shows the probabilities estimated for this example.

The ESL score is the summation of the possible number of non-relevant documents assessed weighted by its probabilities of occurrence as depicted at the last line in Figure 4-13. The browsing path with a lower ESL score is better since the expected number of the non-relevant documents need to be assessed before finding the first relevant document is smaller. Since there are a number of possible browsing paths can be chosen by the user for a given query, the *minimum* ESL (upper bound) and the *average*

ESL will be observed in this evaluation. These values will be used to compare the performance of different implicit RF system in this chapter.

4.4 Experimental Results and Analysis

4.4.1 Experimental setup

The main limitation of the evaluation conducted in (Vinay et al., 2005) is that it was conducted on a small document collection and without proper queries (search tasks). In the evaluation, the query terms were randomly selected from a randomly selected document in the collection and the system was trying to retrieve the selected document based on this artificial query. The aim is to conduct a similar evaluation with a larger test collection and with proper queries.

The ostensive model (Campbell and van Rijsbergen, 1996) is used as the implicit RF model in this experiment. Three ostensive profiles of the ostensive model are evaluated for this IIR system, which are decreasing, increasing and constant profiles (Campbell, 2000). The decreasing profile assumes that the recently viewed documents are more important than the past documents within a browsing path (a browsing session) while the increasing profile assumes otherwise (Campbell, 2000). The constant profile assumes all documents in a browsing path are equally important.

Table 4-1: The description of the implicit RF models

Implicit RF Model	System ID	System Description
Ostensive Model	OMDec	Ostensive Model with decreasing profile
	OMCon	Ostensive Model with constant profile
	OMInc	Ostensive Model with increasing profile

Moreover, the effect of gap in choosing the documents to be displayed to the user from the set of retrieved documents is evaluated. The gap is introduced to provide the user with a set of documents that covers broader topics. In this experiment, performance of the system is evaluated for gap equal to 3, 6 and 9.

4.4.1.1 Research hypotheses

This evaluation is meant to discover the best configuration for the implicit RF system from different implementation models. The main research question is which configuration leads to the optimal performance of the IIR system. The question leads us to ask two related questions, (1) how to choose document candidates for the next

iteration and (2) which document from the candidate set is to be displayed to the user in the next iteration.

The first of the two questions investigates the effectiveness of the implicit RF system for mobile devices compared to a conventional IR system as the baseline. For the baseline, the system displays the top retrieved documents in batches to the user, such as displaying the first top four documents followed by the next top four documents and so on. The implicit RF system dynamically learns the information needs of the user by observing his/her interaction. As a result, more relevant documents should be retrieved by the implicit RF system. Therefore, the implicit RF system should be more effective than the baseline system.

There are three different ostensive profiles used in the model, the decreasing, the constant and the increasing profiles. In (Campbell, 2000), the author claimed that the decreasing profile performs best when he evaluated the ostensive browser for image retrieval. It will be interesting to verify this claim and to observe its performance in the context of larger textual test collections.

Concerning the second of the two questions, the evaluation is focused on the display strategy, the comparison of the top-*k* strategy and the gapped top-*k* strategy. The gapped top-*k* display strategy is meant to create a more diverse set of displayed documents, content-wise. It is believed that by displaying a more diverse set of documents it could lead to better feedback learning for the implicit RF models and to reduce the effect of a search converges to an undesired topic due to an ineffective learning process. Therefore, it is hypothesised that the gapped top-*k* documents strategy is more effective than the top-*k* strategy.

4.4.1.2 Test collections and experimental parameters

For the IIR evaluation, a few parameters need to be set. (1) The first parameter is the number of documents to be displayed each time to the user. In this experiment, the number of displayed documents is set to *four*. It is reasonable considering that the size of display screen for the IIR application being evaluated is small (mobile devices). The same size is chosen by the authors in (Vinay et al., 2005). (2) The second parameter is the number of iterations allowed for the IIR to expand, or the length of browsing. The number of iterations is set to *five* following the authors in (Vinay et al., 2005). (3) The

third parameter is the number of the top weighted index terms to be included in the refined query, *l*, which is set to *ten*. (4) Moreover, for the IR system, the Terrier retrieval system is used with PL2 as the weighting function and the parameter *c* of the Terrier system is not optimized and it is set to the default value, 1.0. Terrier is used as the baseline as well as the retrieval engine for the implicit RF system to ensure its consistency and its independence from the implicit RF models and the display strategy evaluated.

There are three test collections used in this experiment, the TREC 1 ad-hoc retrieval collection (*TREC 1*), the TREC 7 ad-hoc retrieval collection (*TREC 7*) and the TREC 2003 .GOV collection with topic distillation tasks (*TREC .GOV*). Topics 51 to 100 are used for the TREC 1, topics 351 to 400 are used for the TREC 7 and the topic distillation set TD1 – TD50 is used for TREC .GOV.

4.4.2 The effectiveness of the baseline

4.4.2.1 Mean Average Precision (MAP)

As the baseline, the IR system displays the top retrieved documents in batches interactively. For instance, the system returns the first four documents from the ranked list to the user, followed by the next four documents in the next iteration and so on. As a result, the user's browsing path consists of a sequence of documents displayed, four at a time. The effectiveness of the browsing path is measured by the MAP score and the ESL score. Note that, the MAP score is usually calculated for all queries in the collection as a single-value measure for the performance of the system. Throughout the experiments in thesis, the MAP score is reported for each query rather than as an average for all queries in order to investigate the performance of the system in each query.

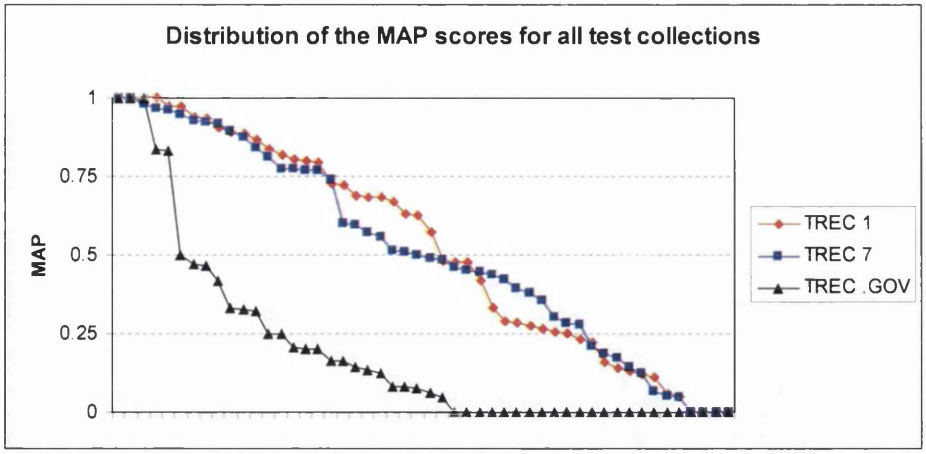


Figure 4-14: The distribution of the MAP scores for all test collections

Figure 4-14 shows the distribution of the MAP scores for the 50 queries of the three test collections. The MAP scores for the TREC 1 and the TREC 7 collections are well distributed among the queries. In the case of the TREC .GOV collection, the MAP scores for the majority of the queries is less than 0.5 while the score of almost half of the 50 queries is zero. Based on the distribution pattern, the queries can be divided into a few categories. In the rest of the experiments, the queries are divided into four categories based on their MAP scores. Table 4-2 shows the ranges of the MAP score for each query category.

Table 4-2: The query category based on the MAP score

Query category	*MAP*
Excellent	0.75 to 1.00
Good	0.50 to 0.74
Moderate	0.25 to 0.49
Poor	0.00 to 0.24

Table 4-3 shows the MAP scores for the baseline system in all test collections. The scores are reported for the excellent, the good, the moderate and the poor queries based on the categories defined in Table 4-2. The average scores for all queries in the test collections are also reported. The number in the brackets indicates the number of queries that belongs to each category.

Table 4-3: The MAP scores for the baseline

Query category	TREC 1	TREC 7	TREC .GOV
Excellent	0.906 (17)	0.890 (17)	0.934 (5)
Good	0.667 (9)	0.574 (8)	0.500 (1)
Moderate	0.345 (11)	0.400 (13)	0.355 (8)
Poor	0.094 (13)	0.084 (12)	0.047 (36)
Overall	**0.528 (50)**	**0.519 (50)**	**0.194 (50)**

Based on Table 4-3, it is obvious that the queries for the TREC 1 and TREC 7 collections are well distributed among the categories. The number of queries fall into each category is almost similar for the TREC 1 and TREC 7 collections. In the case of the TREC .GOV collection, 72% of the queries fall into the poor category with the MAP score for 46% of them is zero. For the TREC 1 and TREC 7 collections, 48% and 50% of the queries are either moderate or poor, respectively, while 88% of the TREC .GOV queries are either moderate or poor. In addition, the overall MAP score for the TREC 1 and TREC 7 collection is almost similar (0.53 and 0.52 respectively), while the overall MAP score for the TREC .GOV collection is low (0.19).

Such a poor performance of the TREC .GOV collection in comparison with the other two collections is due to the size of the collection whereby the TREC .GOV collection is far bigger than the two collections. As a result, the queries may fail to retrieve the relevant document. Note that the MAP score is computed based on the top-20 documents only and not the entire set of retrieved documents. The purpose of choosing only the top-20 documents is to make it easier to compare the performance of the baseline with the performance of the IIR system.

4.4.2.2 Expected Search Length (ESL)

Table 4-4 shows the expected search length measured on the ranking of documents produced by the baseline. Based on the experimental result, some of the queries do not return any relevant documents in its top-*20*. There are 4 queries with no relevant documents retrieved in its top-*20* for the TREC 1 and TREC 7 collections, while the TREC .GOV collection has 23 queries with no relevant documents retrieved.

Table 4-4: The ESL score for the baseline

	TREC 1	TREC 7	TREC .GOV
Average ESL	2.48 (46)	2.48 (46)	5.02 (27)
Number of queries with no relevant document retrieved	4	4	23

The average ESL score is computed based on the remaining queries with at least one relevant documents retrieved. The average ESL for the TREC 1 and TREC 7 collections is 2.48 while the average ESL for the TREC .GOV collection is 5.02. It means that the average expected number of non-relevant documents needs to be assessed before the first relevant document is found is 2.48 and 5.02, respectively. Since there are *four* documents displayed each time in a sequence, the relevant

document can be found in the first sequence of documents displayed for the TREC 1 and TREC 7 collections, while the relevant document can only be found on the second sequence of documents in the case of the TREC .GOV collection. In other words, the user needs to browse to the second *screen* of the result page to find the relevant document for the TREC .GOV collection.

4.4.3 The effectiveness of the implicit RF system

4.4.3.1 Mean Average Precision (MAP)

In the previous section, the performance of the baseline is observed based on the MAP scores computed for each query. It is also learnt that half or more queries fall are either moderate of poor for all three test collections based on their MAP scores. The implicit RF system may improve the performance of the system by adaptively learning from the user feedback to improve the quality of the retrieved documents in a browsing session. In this section, the hypothesis is investigated by evaluating the performance of the ostensive model based implicit RF system for the decreasing, increasing and constant profile.

In the case of the baseline, the MAP scores are computed for a browsing path generated based on the ranking of the top-20 retrieved documents. However, the implicit RF system will generate more than one possible browsing path. In fact there are 256 possible browsing paths generated for each query by limiting the number of documents displayed each time to *four* and the length of browsing to *five*. Each browsing path contains 20 documents.

Due to this scenario, each query will have 256 different MAP scores, one for each browsing path. Therefore, the average MAP scores computed from the 256 browsing paths are compared to the MAP score for the baseline system. This comparison will indicate the average performance of the implicit RF system compares to the baseline system.

Table 4-5: The average MAP scores of the implicit RF system for TREC 1 collection

Query Category	Baseline	OMDec	Δ %	OMCon	Δ %	OMInc	Δ %
Excellent	0.906	0.909	0.28	0.910	0.39	0.914	0.85
Good	0.667	0.684	2.64	0.702	5.28	0.702	5.25
Moderate	0.345	0.351	1.75	0.352	2.16	0.346	0.44
Poor	0.094	0.058	-38.18	0.060	-35.47	0.063	-32.49
Overall	**0.528**	**0.524**	**-0.75**	**0.529**	**0.10**	**0.530**	**0.25**

Table 4-5 shows the MAP scores for the ostensive model based implicit RF system for the TREC 1 test collection. Overall, no significant improvement of MAP scores for the implicit RF systems is observed as compared to the baseline. Only OMCon and OMInc show a very small improvement of 0.1% and 0.25%, respectively, while OMDec is not better than the baseline. A small improvement of less than 1% is observed for the excellent queries for OMDec, OMCon and OMInc. In the case of the good queries, an improvement of more than 5% is observed for OMCon and OMInc, while OMDec shows an improvement of only 2.6%. For moderate queries, the MAP scores for OMDec and OMCon are 2% better than the baseline, while less than 0.5% improvement is shown for OMInc. One interesting observation is that none of the implicit RF system improves the MAP scores for the poor queries. The difference is ranges from 32.5% to 38.2% less than the baseline.

Based on Table 4-5, for the TREC 1 collection, the average performance of the browsing paths generated by the implicit RF system for each query is not significantly better than the performance of the baseline for the query. Moreover, the performance of OMDec (decreasing profile), OMCon (constant profile) and OMInc (increasing profile) is almost similar and OMInc is the best among them for this test collection.

Next, the performance of the implicit RF system in the TREC 7 collection is observed. Table 4-6 shows the MAP scores of the implicit RF system for the TREC 7 collection based on the query categories and also for the overall queries. Overall, a significant improvement of the MAP scores for the implicit RF system is observed as compared to the baseline. The improvement of 6.57% (p=0.049, paired t-test), 8.39% (p=0.021, paired t-test) and 8.78% (p=0.011, paired t-test) for OMDec, OMCon and OMInc, respectively is significant, indicated by (*) symbol in Table 4-6. Similar to the TREC 1 collection, OMInc appears to be the best model for the TREC 7 collection.

Table 4-6: The average MAP scores of the implicit RF system for TREC 7 collection

Query Category	Baseline	OMDec	Δ %	OMCon	Δ %	OMInc	Δ %
Excellent	0.890	0.898	0.92	0.909	2.15	0.913	2.62
Good	0.574	0.710	23.70	0.749	30.45	0.750	30.66
Moderate	0.400	0.457	14.21	0.457	14.43	0.460	14.99
Poor	0.084	0.062	-26.26	0.059	-29.84	0.058	-30.60
Overall	**0.519**	**0.553**	**6.57***	**0.562**	**8.39***	**0.564**	**8.78***

In the case of the excellent queries, a small improvement ranging from 0.9% to 2.6% is observed for the implicit RF system. For the good queries, the MAP scores for the implicit RF system are around 23.7% to 30.7% better than the baseline. An about 14% to 15% improvement can be observed for the moderate queries. Similar to the TREC 1 collection, the performance of the poor queries reduces with 26% to 30% less than the baseline. Such an improvement depicted in Table 4-6 shows that the implicit RF system has the potential to improve the performance of the baseline.

Table 4-7: The average MAP scores of the implicit RF system for TREC .GOV collection

Query Category	Baseline	OMDec	Δ %	OMCon	Δ %	OMInc	Δ %
Excellent	0.934	0.374	-59.95	0.385	-58.72	0.392	-58.06
Good	0.500	0.000	-100.00	0.000	-100.00	0.000	-100.00
Moderate	0.355	0.197	-44.33	0.203	-42.84	0.204	-42.40
Poor	0.047	0.161	242.08*	0.158	237.22*	0.157	234.31*
Overall	**0.194**	**0.185**	**-4.76**	**0.185**	**-4.58**	**0.185**	**-4.64**

In the case of the TREC .GOV collection, a slightly different pattern is depicted in Table 4-7. All queries show no improvement in the MAP scores except the poor queries. An improvement of 242% (p=0.029, paired t-test), 237% (p=0.032, paired t-test) and 234% (p=0.035, paired t-test) for OMDec, OMCon and OMInc, respectively, is significant. However, the average MAP scores for all queries are about 5% less than the baseline. Moreover, the performance of OMDec, OMCon and OMInc is almost similar.

Based on the results of comparing the average MAP scores computed based on the possible browsing paths generated by the system to the MAP score of the baseline, it is not clear the benefit of the implicit RF model in improving the effectiveness of the retrieval system. A significant improvement can be observed only for the TREC 7 collection and for the poor queries of the TREC .GOV collection. It is interesting to find out the queries that improve the performance of the system. Table 4-8 shows the percentage of the queries that improve the MAP scores of the baseline. Based on the table, the improvement in the MAP scores occurs in more than half of the queries in the TREC 7 collection. In the case of the other test collections, the improvement of the MAP scores is observed for slightly less than half of the queries. Based on this result, it is learnt that the implicit RF model has a good potential in improving the effectiveness of the retrieval system.

Table 4-8: The percentage of queries that improve the baseline

	OMDec	OMCon	OMInc
TREC 1	44%	52%	46%
TREC 7	56%	62%	62%
TREC .GOV	44%	46%	46%

So far, the analysis is conducted based on the average MAP score computed for the possible browsing paths generated by the implicit RF system. The average MAP score is compared to the MAP score of the baseline. Since the implicit RF system offers a set of possible browsing paths rather than a single browsing path, the average MAP score is used to indicate the average performance of the system.

Table 4-9: The average chances of improving the MAP scores of the baseline

	OMDec	OMCon	OMInc
TREC 1	35.78%	37.32%	39.49%
TREC 7	52.30%	55.29%	56.63%
TREC .GOV	32.98%	33.08%	32.02%

It is also interesting to look into the chances that the user chooses a browsing path with a better MAP score than the baseline. The chances can be computed as a ratio of the number of browsing paths that improve the scores to the total number of possible browsing paths generated by the system for each query. Table 4-9 shows the average chances of improving the MAP scores for all queries in the test collections. There are more than 50% chances of improving the MAP scores of the baseline for the TREC 7 collection by using the implicit RF system. OMInc appears to be the best model among them. The chances for the TREC 1 collection is in the range of 35.8% to 39.5% while the chances for the TREC .GOV is about 32% to 33%.

Up to this stage of discussion, it is learnt that the use of the implicit RF model may improve the effectiveness of the baseline system, especially for the TREC 7 collection. Furthermore, the effectiveness of about half of the queries in each test collection can be improved. However, the chances of choosing the browsing path that is better than the baseline is rather small with about 32% to 57%. Next, the maximum MAP score that can be achieved by the implicit RF system for each query is investigated to observe the upper limit performance of the system. It is the highest MAP score computed from the set of possible browsing paths generated for the query.

Table 4-10: The maximum MAP scores of the implicit RF system for TREC 1 collection

Query Category	Baseline	OMDec	Δ %	OMCon	Δ %	OMInc	Δ %
Excellent	0.906	0.959	5.80	0.964	6.31	0.964	6.33
Good	0.667	0.803	20.46	0.813	22.05	0.815	22.24
Moderate	0.345	0.474	37.60	0.483	40.09	0.473	37.13
Poor	0.094	0.177	89.50	0.197	110.76	0.209	122.86
Overall	**0.528**	**0.621**	**17.55**	**0.632**	**19.55**	**0.633**	**19.73**

Table 4-10 shows the maximum MAP scores of the implicit RF system for the TREC 1 collection. Based on the table, it is learnt that the maximum improvement of the MAP score is about 17.6% to 19.7% over the baseline for all queries in the test collection. Moreover, the performance of the system for the queries in all categories can be improved and the highest improvement can occur for the poor queries. A similar pattern can be observed for the TREC 7 collection as depicted in Table 4-11. The maximum improvement of 24.3% to 26.5% can be observed for all queries in the test collection. Moreover, the highest improvement can occur for the poor queries.

Table 4-11: The maximum MAP scores of the implicit RF system for TREC 7 collection

Query Category	Baseline	OMDec	Δ %	OMCon	Δ %	OMInc	Δ %
Excellent	0.890	0.962	8.09	0.968	8.77	0.970	8.95
Good	0.574	0.841	46.59	0.841	46.59	0.859	49.71
Moderate	0.400	0.553	38.31	0.560	40.13	0.564	41.10
Poor	0.084	0.163	94.40	0.168	100.14	0.175	109.07
Overall	**0.519**	**0.645**	**24.31**	**0.650**	**25.30**	**0.656**	**26.50**

For the TREC .GOV collection, a different pattern can be observed. The maximum improvement of 25.7% to 29.4% of the MAP scores can be observed for all queries. However, the improvement is only applicable for the poor queries as the maximum MAP score for the other queries are less than the MAP score of the baseline. In the case of the poor queries, the maximum improvement is about 346% to 360% over the baseline.

Table 4-12: The maximum MAP scores of the implicit RF system for TREC .GOV collection

Query Category	Baseline	OMDec	Δ %	OMCon	Δ %	OMInc	Δ %
Excellent	0.934	0.497	-46.72	0.499	-46.58	0.499	-46.53
Good	0.500	0.0	-100.00	0.000	-100.00	0.000	-100.00
Moderate	0.355	0.270	-23.90	0.281	-20.78	0.285	-19.68
Poor	0.047	0.209	345.64	0.212	352.00	0.216	359.67
Overall	**0.194**	**0.244**	**25.65**	**0.248**	**27.74**	**0.251**	**29.42**

Based on the results discussed so far, it is learnt that the implicit RF system could improve the effectiveness of the retrieval system. In this experiment, the same retrieval

engine, the Terrier system with the same configuration, is used for the baseline and also for the implicit RF system. The only difference is the existence of the implicit RF model namely the ostensive model with three different profiles (Campbell and van Rijsbergen, 1996) to learn the user information need based on his/her feedback.

However, due to the fact that there is more than one possible browsing path that the user can select, the potential benefit of the implicit RF system is not obvious. Based on the results, the chances that the user chooses a better browsing path is about 32% to 39% for the TREC 1 and TREC .GOV collections, and is about slightly more than 50% for the TREC 7 collection. Moreover, there are plenty of space for improvement can be observed based on the upper limit improvement presented earlier. As a result, the system should provide a browsing recommendation capability to ensure that the user chooses the right browsing path. In this thesis, such a recommendation capability can be achieved by incorporating the intertemporal choice browsing model into the implicit RF system, which will be discussed in the next chapter.

4.4.3.2 Expected Search Length (ESL)

In Section 4.4.2.2, the performance of the baseline is measured based on the ESL score. Based on the results, the queries of the test collection are divided into two groups, the queries with at least one relevant document retrieved in its top-*20* and the queries with no relevant document retrieved. In this section, the minimum ESL (upper-bound performance) of the implicit RF model is compared to the ESL of the baseline to observe the ranges of improvement (if any) that may be produced by the implicit RF models over the baseline. Table 4-13 shows the minimum ESL score of the implicit RF models for the queries with at least one relevant document retrieved in the top-*20* and Table 4-14 shows the minimum ESL score of the implicit RF models for queries with no relevant document retrieved.

Table 4-13: The minimum ESL for queries with at least one relevant document

	Baseline	OMDec	Δ %	OMCon	Δ %	OMInc	Δ %
TREC 1	2.48	2.61	-5.12	2.68	-8.26	2.61	-5.48
TREC 7	2.42	2.17	10.40	2.33	3.96	2.26	6.81
TREC .GOV	5.02	4.94	1.54	4.82	3.87	5.27	-5.10
Average			**2.27**		**-0.14**		**-1.26**

Based on Table 4-13, it is obvious that, on average, all the implicit RF models will not improve the performance of the baseline for the TREC 1 collection. Meanwhile, a maximum improvement of 10.4%, 3.96% and 6.81% can be expected from OMDec,

OMCon and OMInc, respectively, for the TREC 7 collection if the user chooses the right browsing path. In the case of the TREC .GOV collection, a maximum improvement of 1.5% and 3.9% can be expected from OMDec and OMCon, respectively, OMInc is unable to improve the ESL of the baseline.

Table 4-14: The minimum ESL for queries with no relevant document retrieved

	Baseline	OMDec	Δ %	OMCon	Δ %	OMInc	Δ %
TREC 1	20+	14.54	>27.29	13.54	>32.29	13.33	>33.33
TREC 7	20+	12.13	>39.38	10.13	>49.38	11.13	>44.38
TREC .GOV	20+	15.05	>24.75	14.89	>25.56	14.15	>29.26
Average			**>30.47**		**>35.74**		**>35.66**

In Table 4-14, it can be observed that the implicit RF models may tremendously improve the performance of the baseline in the context of ESL for the queries with no relevant document retrieved by the baseline in its top-*20*. Since there is no relevant document, the ESL is more than 20 as indicated by (+) symbol in the table. Based on the table, it is learnt that the maximum improvement of the ESL score is ranging from 27% to 33% for the TREC 1 collection, from 39% to 49% for the TREC 7 collection and from 25% to 29% for the TREC .GOV collection. The results show that the implicit RF models have a potential to retrieve relevant documents for those queries.

4.4.4 The effect of gap of the display strategy on the MAP

In Section 4.2.1, the display strategy is described as one of the important components of the implicit RF system for mobile devices. The aim of the display strategy is to choose the documents to be displayed such that the user will quickly find the relevant documents as well as the system will be able to effectively learn the user feedback from his/her interaction with the displayed documents. In the earlier experiments, the top-k documents are displayed to the user, where k is equal to 4. However, the strategy may not be effective for the implicit RF model (refer to Section 4.2.1). In this section, the performance of the implicit RF system is evaluated when the Gapped Top-k display strategy is used. In particular, the effect of the increasing size of the gap on the MAP score of the system is tested.

Table 4-15: The MAP score of OMDec as the gap size increases for TREC 1

Query Category	No Gap	Gap = 3	Δ %	Gap = 6	Δ %	Gap = 9	Δ %
Excellent	0.909	0.903	-0.66	0.866	-4.67	0.831	-8.54
Good	0.684	0.762	11.39	0.738	7.82	0.749	9.47
Moderate	0.351	0.389	10.89	0.382	8.91	0.323	-7.86
Poor	0.058	0.097	66.82	0.115	98.29	0.120	106.72
Overall	**0.524**	**0.555**	**5.81**	**0.541**	**3.22**	**0.520**	**-0.91**

Table 4-16: The MAP score of OMDec as the gap size increases for TREC 7

Query Category	No Gap	Gap = 3	Δ %	Gap = 6	Δ %	Gap = 9	Δ %
Excellent	0.898	0.838	-6.75	0.826	-8.10	0.818	-8.93
Good	0.710	0.795	11.90	0.754	6.19	0.723	1.79
Moderate	0.457	0.278	-39.07	0.302	-33.78	0.291	-36.37
Poor	0.062	0.150	142.55	0.130	110.00	0.063	1.20
Overall	**0.553**	**0.520**	**-5.85**	**0.511**	**-7.51**	**0.484**	**-12.35**

Table 4-17: The MAP score of OMDec as the gap size increases for TREC .GOV

Query Category	No Gap	Gap = 3	Δ %	Gap = 6	Δ %	Gap = 9	Δ %
Excellent	0.374	0.385	2.99	0.327	-12.43	0.394	5.34
Good	0.000	0.000	0.00	0.000	0.00	0.000	0.00
Moderate	0.197	0.236	19.41	0.222	12.39	0.201	1.63
Poor	0.161	0.157	-2.18	0.139	-13.80	0.158	-1.42
Overall	**0.185**	**0.189**	**2.60**	**0.168**	**-9.04**	**0.186**	**0.47**

The use of the gap for the display strategy does not give a significant improvement to the performance of the implicit RF system as depicted in Table 4-15, Table 4-16 and Table 4-17. A small improvement of 5% or less can be observed in few cases, while the other cases show that the performance of the system decreases. Moreover, as the size of the gap increases, the performance of the implicit RF system is not necessarily increases. Most likely, the performance will decrease as shown in Figure 4-15.

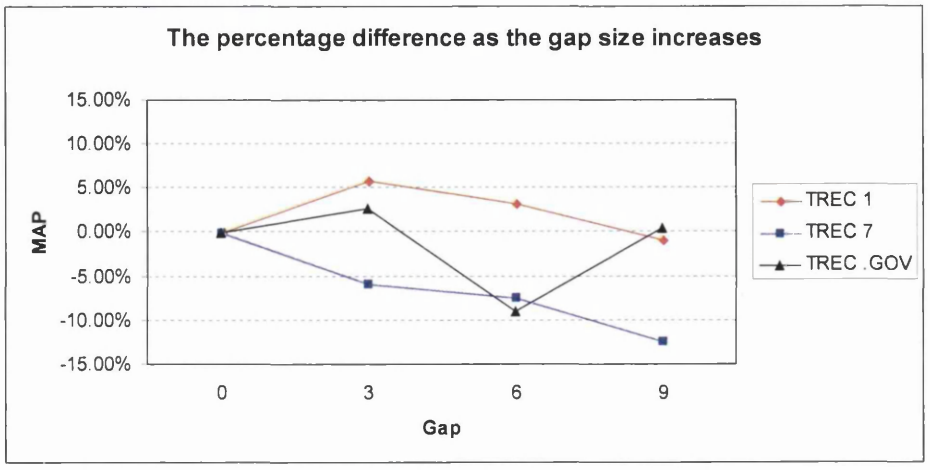


Figure 4-15: The percentage difference of the MAP score as the gap size increases

A higher MAP score means that more relevant documents can be found at the top of the ranking and *vice versa*. Therefore, the use gap may not be useful for the excellent queries and may be very useful for the poor queries. Based on Table 4-15, Table 4-16 and Table 4-17, it is obvious that the excellent queries do not benefit from the use of the gap in the display strategy. The MAP scores for the excellent queries reduces as the gap size increases for the TREC 1 and TREC 7 collections, while the MAP scores may slightly improve when gap 3 and 9 is used for the TREC .GOV collection.

In the case of the poor queries, the MAP scores for the TREC 1 collection reduces as the gap size increases. However, the improvement of the MAP scores for the poor queries of the TREC 7 collection is not may not depend on the size of the gap. In the case of the TREC .GOV collection, no improvement of the MAP scores can be observed for different gap sizes. Based on these results, it is learnt that the gap size affect the performance of the system but no obvious pattern can be found. Moreover, it is difficult to determine the optimal size of the gap based on the findings so far.

4.5 Discussion

In this chapter, the effectiveness of the implicit relevance feedback (RF) models for mobile devices has been investigated. The aim is to overcome the two main limitations of mobile devices, its small display unit and its limited interaction capability. The problem with the small screen can be overcome by displaying fewer retrieved documents chosen by the display strategy module to optimise the performance of the system. Meanwhile, the problem of limited interaction capability is overcome by adopting an implicit relevance feedback model to infer the user information need from his/her interaction. Based on the evaluation results, the implicit RF system has a good potential to improve the effectiveness of the IR system for mobile devices.

4.5.1 The effectiveness of the implicit RF system for mobile devices

Relevance feedback (RF) has become one of the most important techniques to improve the effectiveness of an information retrieval (IR) system (Salton and Buckley, 1990). The technique aims to refine a query such that it will be much closer to the relevant documents. Therefore, more relevant documents can be retrieved. However, the main problem with the RF technique is the need for the user to provide explicit relevance judgements to the system. For instance, the user is required to explicitly indicate which

documents are relevant. The pseudo-relevance feedback technique overcomes the problem by treating the top ranked documents as relevant (Tao and Zhai, 2006). Therefore, the need to explicitly provide relevance information to the system is eliminated. Alternatively, the implicit relevance feedback technique, usually applied in an interactive IR environment, assumes that the documents or the information objects[3] chosen by the user while interacting with the system are relevant (Campbell and van Rijsbergen, 1996, Kelly and Teevan, 2003, Vinay et al., 2005, White et al., 2005) .

In this chapter, the effectiveness of the ostensive model as the implicit RF model is investigated in the context of retrieval system for mobile devices. The effectiveness of the models is compared against the baseline, which is a conventional top ranking system. However, a significant improvement of the system's effectiveness (based on the MAP score) for all queries can be observed only on one test collection, the TREC 7 collection. In addition, a significant improvement of effectiveness can also be observed on the poor queries of the TREC .GOV collection, while no improvement is observed for the poor queries in the other test collections. Such an observation has been discussed in Section 4.4.3.1.

The results of the evaluation are somewhat disappointing as the implicit RF system is not necessarily more effective, on average, than the baseline system. In (Vinay et al., 2005), the authors discovered a similar result, as their relevance feedback models are unable to show a positive improvement. They claimed that the lack of evidence to be used for learning is probably the reason why the implicit RF system fails to improve the effectiveness of the system, since only one document per iteration is indicated as relevant (selected by the user). It is believed that a similar argument applies in this evaluation. Moreover, the evidence from the implicit RF model is implied rather than accurately indicated by the user. In the case of the system proposed in (White et al., 2005), there is more evidence available since it is gathered from a number of different information representations, such as the sentences and the summaries.

On a positive side, the implicit RF system has a potential to improve the effectiveness of browsing as compared to the baseline system. According to the experimental results in Section 4.4.3.1, the average maximum MAP score that the implicit RF system is

[3] The information object refers to all forms of information such as a document, a paragraph, a sentence, a summary and so on.

capable to produce is as high as 20% more than the baseline for the TREC 1 collection, 27% more than the baseline for the TREC 7 collection and 29% more than the baseline for the TREC .GOV collection. Such an improvement can only be enjoyed if the user is capable of choosing the right browsing path from the set of possible browsing paths provided by the implicit RF system. The browsing recommendation model for the implicit RF system to suggest a good browsing path to the user will be discussed in the next chapter.

In this chapter, the cost of browsing to find the first relevant document in a given browsing path (session) is measured based on the expected search length (ESL) (Cooper, 1968). It measures the expected number of non-relevant documents needed to be assessed before finding a given number of relevant documents. Based on the results in Section 4.4.3.2, it appears that there is a potential to minimise the cost of finding the first relevant document for the TREC 7 collection and also in some cases for the TREC .GOV collection. Moreover, for those unsuccessful queries of the baseline system (the queries with no relevant document retrieved at top-*20*), the cost of browsing can be minimized due to the fact that the implicit RF system manages to retrieve at least one relevant document for some of those queries.

To summarise, the implicit RF system investigated in this chapter could be a better solution for IR system on mobile devices. However, there is a need to include a browsing recommendation capability in the system to ensure that the user can actually benefit from the RF system.

4.5.2 The effectiveness of the ostensive model for the implicit RF system

In this chapter, the effectiveness of the ostensive model for the implicit RF system is investigated. The model was proposed as a relevance feedback model for the *Ostensive Browser* system (Campbell and van Rijsbergen, 1996, Campbell, 2000, Urban et al., 2003). The browser is used in the context of image retrieval. The experimental results showed that the ostensive model could be a more effective way to retrieve relevant images. In (White et al., 2005), the authors showed that the ostensive model could also be applied in textual information retrieval. Based on the evaluation discussed in this chapter, the ostensive model shows a promising result as a relevance feedback model for an implicit RF system.

There are three different ostensive profiles suggested in (Campbell and van Rijsbergen, 1996, Campbell, 2000), which are the decreasing, the constant and the increasing profile. The decreasing, the increasing and the constant profiles assume that the weight of a document decreases, increases and remains constant as the age of that document in the browsing path increases, respectively. Based on the evaluation in (Campbell, 2000), the decreasing profile is better than the other profiles. Moreover, in (Hirashima et al., 1998), the author suggested that a similar decreasing function is the best . In this evaluation, the increasing profile was shown to be slightly better than the other profiles in the context of an implicit RF system for mobile devices. However, the difference is insignificant. The results were measured based on the average MAP score of the implicit RF system. Based on the experimental results in this chapter, it is not possible to determine the best ostensive profile among them.

4.5.3 The effect of the gap for the display strategy

The introduction of the gap in selecting documents to be displayed to the user is meant to provide a set of displayed documents with a broader topic. The reason is to avoid the condition, where the search or the browse converges to a specific topic that is not intended by the user (Vinay et al., 2005). Instead, a set of documents with a broader topic allows the user to choose the relevant subtopic(s) that he/she is interested in. Also, the effectiveness of the relevance feedback model depends on the documents chosen by the user, thus it depends on the documents displayed to the user (Shen and Zhai, 2005).

Based on the experimental results in Section 4.4.4, it appears that a bigger gap size does not necessarily improve the effectiveness of the implicit RF system. It seems that the MAP score of the system may decreases as the gap size increases.

4.6 Chapter Summary

In this chapter, the effectiveness of an implicit RF system for mobile devices is investigated. The effectiveness of the ostensive model as an implicit RF system is also evaluated. In addition, the effect of display strategy for the performance of the system is observed. Based on the investigation, the implicit RF system could be an effective solution for IR system on mobile devices. However, based on the evaluation results, the real benefit of the system can only be realised if an effective browsing recommendation

model is incorporated to suggest the best browsing path for the user. In the next chapter (Chapter 5), a browsing recommendation model is proposed based on the intertemporal choice browsing model.

5 The Browsing Recommendation Model for the Implicit RF System

5.1 Introduction

In Chapter 4, the implicit relevance feedback (RF) system as an effective information retrieval (IR) system for mobile devices was presented. The system was first proposed in (Vinay et al., 2005) as a more effective algorithm for the retrieval problem. In the last chapter, the use of the ostensive model (Campbell and van Rijsbergen, 1996, Campbell, 2000, Urban et al., 2003) as an implicit RF model for the system was investigated. The results showed that the implicit RF system has a potential to improve the effectiveness of browsing. However, the effectiveness of the implicit RF system is very much depending on the user making the right decision during the browsing session. Therefore, the aim of this chapter is to produce an effective method to choose the best browsing path for the user to browse.

In Chapter 3, an intertemporal choice browsing model as a decision model for selecting the best browsing path for the user was proposed. By treating the problem of browsing as an intertemporal choice problem, the best browsing path can be discovered by using the intertemporal choice model (Read, 2004). Due to the fact that the model has been successful in economic and social choice studies (Samuelson, 1937, Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Lazaro et al., 2002), it is assumed that it can model the decision behaviour of the user in browsing. Thus, an effective browsing recommendation model can be implemented based on the intertemporal choice browsing model.

Based on the assumption above, the intertemporal choice browsing model is adopted to select the best browsing path for the implicit RF system discussed in Chapter 4. In this chapter, the issues regarding the implementation of the intertemporal choice browsing model as a browsing recommendation model for the implicit RF system will be discussed. An investigation into the effectiveness of the intertemporal choice browsing model to recommend a browsing path to the user will be conducted. The aim of this chapter is to show that the effectiveness of an implicit RF system can be improved by incorporating the intertemporal choice browsing model as the browsing recommendation model.

In Section 5.2, a browsing recommendation model that is based on the intertemporal choice browsing model proposed in Chapter 3 is presented. Next, in Section 5.3, the methodology to evaluate the browsing recommendation model in the context of the implicit RF system is discussed. Then, the results of the experiments are presented and analysed in Section 5.4. After that, the discussion on the results of the experiments is presented in Section 5.5. Finally, in Section 5.6, a summary for this chapter is provided.

5.2 The Browsing Recommendation Model

The purpose of the browsing recommendation model is to select the best browsing path from a set of all possible browsing paths. The selected browsing path will be recommended to the user to help him/her during browsing. The need for such a model arises due to the limitation of the implicit RF system discussed in Chapter 4. Due to the limitation, there is no guarantee that the user can benefit from the system since there are many ways to browse the system and any given browsing path will not necessarily lead to a better browsing session. As a result, a browsing recommendation model is required to provide some guidance to the user.

There are two important aspects of the browsing recommendation model, the understanding of the user's need and the subsequent suggestion of documents to be browsed. In that sense, any implicit RF system is also a recommendation system. The query refinement process in the implicit RF system is a way to better understand the information need of a user which leads to the retrieval of a better set of documents to satisfy the need. Many implicit RF systems adopt this mechanism (Campbell and van Rijsbergen, 1996, Campbell, 2000, Urban et al., 2003, White et al., 2004, Vinay et al., 2005). All of these systems share the same technique where the user's information need

is implied through the documents selected by the user and subsequently the potentially relevant documents are returned to the user. In the Web domain, the information need of the user is learnt from the hyperlinks selected while browsing. Subsequently, recommended hyperlinks or documents are displayed to help the user to choose the next hyperlinks or documents (Armstrong et al., 1995, Lieberman, 1995).

Alternatively, a collaborative recommendation technique can be adopted, which is based on analysing frequent browsing paths chosen by other users to suggest documents to the user. This technique is usually implemented in the Web environment since frequent browsing paths can be easily gathered from Web access logs[4]. A number of Web usage mining methods (Dunham, 2003) can be applied to the data from the logs to construct a knowledge base for the system. The recommendation technique observes the documents already selected by the user in the browsing session and subsequently suggests other unseen documents in the collection chosen by other users with similar documents selection. For instance, the *Footprint* system shows a map of other users' frequent browsing paths to help the user choose the next document to browse (Wexelblat and Maes, 1997). The recommended documents or hyperlinks can be shown to the user as he/she browses the Web by using a clustering-based recommendation (Mobasher et al., 2002) or association rules-based recommendation (Mobasher et al., 2001). The browsing session of the user can also be improved by dynamically improving the structure of the Web through clustering the log data (Perkowitz and Etzioni, 2000). The combination of content-based recommendation based on information need discovery through IR techniques and usage-based recommendation based on mining Web access logs could improve the performance of the recommendation system (Azman and Ounis, 2004).

However, neither method may produce an optimal recommendation for the user. The major problem of current recommendation techniques is that recommendations are computed individually for each step of the browsing session. As we have seen before, the implicit RF system is unable to guarantee optimal browsing for the user, based on the results from the experiments conducted in Chapter 4. The solution to the problem will be to choose one browsing path from all possible browsing paths of the user. In order to do that, the system should be able to measure the usefulness of a browsing path and should be able to assign an appropriate score for the browsing path. The best

[4] Web access logs contain the information about the documents that the users visited.

browsing path could easily be chosen based on the one with the highest score. The browsing recommendation model discussed in this chapter is based on this idea.

5.2.1 The Recommendation Model

In order to recommend a browsing path (or a browsing strategy), the model should be able to discover all the possible browsing paths that can be chosen by the user. In the case of the implicit RF system, the browsing paths can be discovered by observing all possible user interactions while browsing. All possible interactions between the user and the system can be represented as a decision tree such as the one depicted in Figure 4-9. The decision tree also shows the behaviour of the system responding to all possible interactions of the user.

Based on the decision tree, a set of possible browsing paths can be generated and the user will choose only one of these browsing paths. For instance, if the system displays *four* documents each time, there are four possible ways to choose the documents. If the length of the browsing session is set to *five*, which means that there will be five iterations and there will be $4 \times 4 \times 4 \times 4 = 256$ possible browsing paths. Note that, a browsing path is a sequence of documents displayed to the user, where in each sequence consists of a set of documents such as the one depicted in Figure 4-10.

Given that $path$ is the set of possible browsing paths (that can be extracted from a decision tree) and $path_k \in path$ is one of the browsing paths. Let U_k be the utility or the score for the browsing path $path_k$, the best browsing path $path_k^*$ is the one with the highest score, such that $U_k^* = \max\{U_k\}$.

The score of a browsing path, U_k, is computed based on the intertemporal choice browsing model. Let $U(u_1, t_1; ...; u_n, t_n)$ be the sequence of values for the browsing path $path_k$, the utility or the score for the browsing path, U_k can be estimated based on the following equation.

$$U_k = U(u_1, t_1; \ldots; u_n, t_n) = \sum_{i=1}^{n} v(u_i)\phi(t_i) \quad (5\text{-}1)$$

where $v(u)$ is the value function and $\phi(t)$ is the discount function. A detailed discussion on the intertemporal choice browsing model has been provided in Chapter 3.

5.2.2 Constructing a browsing path

One of the important issues regarding the application of the intertemporal choice browsing model is the formulation of the sequence of values from a browsing path. According to Figure 5-1, a browsing path is a sequence of *screens* displayed to the user as a result of the user interactions where each screen consists of four documents. The value of each screen to the user, such as the usefulness of a particular screen in the context of the user information seeking session, depends on the value of the documents displayed on the screen. Therefore, the value of a browsing path to the user depends on the value of the screens iteratively displayed to him/her. As such, a browsing path is assumed to be a sequence of screen values.

The screen's value should be derived from the value of the documents contained in the screen. Therefore, there are four possible values for each screen, one from each document contained in the screen. Hence, the problem of constructing a browsing path is reduced to the problem of choosing which value of the documents to be used as the value of the screen. In this chapter, three strategies are investigated to choose the value for the screen, the *SingleValue*, the *EqualWeight* and the *DecreasingWeight*.

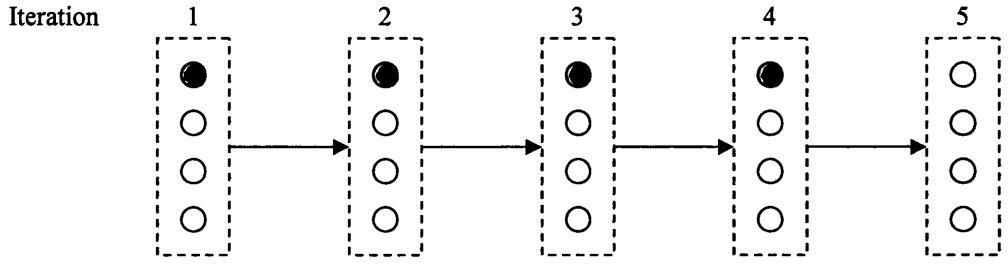


Figure 5-1: The browsing path generated by always selecting the top document

Figure 5-1 shows an example of a browsing path generated when the user always chooses the top document displayed to him/her. The blue-coloured circles represent the selected documents.

First, the *SingleValue* strategy assumes that the value of each screen can be represented by the value of one of the documents. In particular, the strategy suggests that the value of each screen should be represented by the value of the selected document. It is based on the assumption that the selected document in each screen is most relevant to the user's information need (Campbell and van Rijsbergen, 1996, White et al., 2004, Vinay et al., 2005). As such, the value of each screen can be represented by the chosen document. Therefore, a browsing path is represented by a sequence of the values for the selected documents. Based on Figure 5-1, a browsing path is represented by the sequence of values of the blue-coloured circles.

Second, the *EqualWeight* and the *DecreasingWeight* techniques assume that the value of each screen should be represented by the values of all documents displayed on the screen. It is based on the assumption that the effectiveness of a browsing path depends on all the documents displayed on each screen iteratively. Since there is more than one possible value for each screen (or for each state of browsing), the intertemporal choice browsing model for uncertain outcomes is more suitable for this scenario (refer to Section 3.3.3.4 for a detailed description of the model). In short, each document in the screen is assigned with a probabilistic weight corresponding to the importance of the document to the user. Without prior knowledge about the nature of the weight to be estimated for the documents, an equal weight can be assigned to each document, which is referred to the *EqualWeight* technique.

Alternatively, the documents on the screen can be assigned with different weights based on the probability of the documents being relevant. The documents are usually ranked such that the document at higher rank is more relevant to the user's information need (Robertson, 1977, Gordon and Lenk, 1991). As such, the weight for a document decreases as the rank position of the document increases. As a result, a simple decreasing function can be used to estimate the weight of a document based on its rank on the screen. This technique is referred as the *DecreasingWeight*.

For the *EqualWeight* and the *DecreasingWeight* techniques, the value of each screen is given by the expected value for the documents on the screen. It is a weighted sum of the value of the documents, where the probabilistic weight for each document is assigned based on the assumption of the *EqualWeight* and the *DecreasingWeight*

techniques. In the case of the *EqualWeight* technique, each screen is assigned the average value for the documents.

In the next section, a methodology to evaluate the browsing recommendation model for an implicit RF system is described.

5.3 Evaluation Methodology

The experiments conducted in this chapter are a continuation of the evaluation in Chapter 4. Figure 5-2 shows an implicit RF system after the incorporation of the browsing recommendation model into the system architecture. The role of the browsing recommendation model (*BM*) is to choose the best browsing path (BP) from a set of possible browsing paths extracted from the decision tree (*DT*). The length of the browsing paths and the number of documents displayed by the implicit RF system is fixed in order to generate the decision tree. An example of the decision tree is depicted in Figure 4-9.

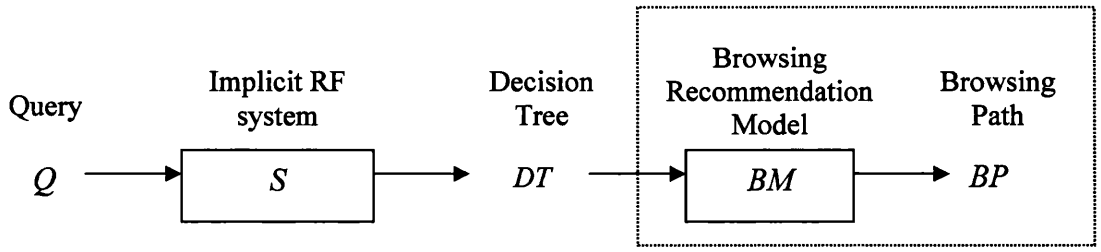


Figure 5-2: The browsing recommendation model incorporated into the implicit RF system

As can be seen in Figure 5-2, for each query the implicit RF system produces a decision tree, while the browsing recommendation model selects a single browsing path from this tree. The decision tree was used to evaluate the effectiveness of the implicit RF system as investigated in Chapter 4. The effectiveness of the implicit RF system is measured based on its average MAP computed from the set of possible browsing paths. The lowest expected search length (ESL) is also noted. In this evaluation, the performance of the browsing recommendation model is measured based on the effectiveness of the selected browsing path (*BP*). The effectiveness of the browsing path is measured based on its MAP score and also its ESL score. The discussion regarding the evaluation measures are described in Section 4.3.1.

5.3.1 The setup for the intertemporal choice model

The browsing recommendation model proposed in this research is based on the *intertemporal choice* model (Samuelson, 1937, Loewenstein and Prelec, 1992, Frederick et al., 2002, Read, 2004). The two main components of the model are the *value function* and the *discount function*. The value function transforms a value of an outcome into a utility. A monotonic concave increasing function for a positive value and a monotonic convex decreasing function for a negative value has been well accepted as the transformation function (Kahneman and Tversky, 1979) and will be used in this model.

Table 5-1: The value function and its parameter

	The function	*Parameter*
Value function	$v(u) = \log_{10}(au+1)$	a

For this evaluation, it is assumed that the value of the document is non-negative. Therefore, the function in Table 5-1 will be used as the value function. The value of the parameter, a, is set to 1.0. The effect of different values for this parameter is not investigated in this experiment. The main reason is that the utility or the subjective value of a document computed by the value function can only be verified by the user and the user is not present in this simulation based evaluation.

Another issue regarding the value function is the assumption used to estimate the value of a document. The value of a document is estimated based on the query submitted by the user. This assumption is generally accepted in IR research and many IR models are built based on this assumption. It is called the *independent value* of a document. In this evaluation, the independent value of a document is estimated based on relevance value or the retrieval status value (RSV) computed by the IR system.

In the case of the discount function, there are two types of discounting being discussed in the intertemporal choice model, the *positive time preference* and the *negative time preference* (Loewenstein and Prelec, 1993). The negative time preference will not be investigated here since the evaluation model is built based on the assumption of positive time preference for browsing. There are two shapes of discount functions being proposed, an *exponential* discount function and a *hyperbolic* discount function (Frederick et al., 2002). An exponential discounting assumes that the discounting rate is constant as the delay increases while a hyperbolic discounting assumes that the

discounting rate is decreasing as the delay increases. Researchers have found that the hyperbolic discount function is more appropriate for the intertemporal choice problem (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Frederick et al., 2002, Lazaro et al., 2002). This evaluation will attempt to verify this hypothesis for an IR application, particularly in the context of the implicit RF system that is being investigated. The discount functions being investigated in this evaluation are summarised in Table 5-2. The discussion on these discount functions has been provided in Section 3.2.1.2.

Table 5-2: The mathematical equations for the discount functions used in the evaluation

Shape	*Function ID*	*Equation*	*Parameter*
Exponential	SAMUELSON	$\phi(t) = \frac{1}{(1+r)^t}$	r
Hyperbolic	MAZUR	$\phi(t) = \frac{1}{(1+st)}$	s
	HARVEY	$\phi(t) = \frac{1}{(1+t)^h}$	h

Note that, the discount functions in Table 5-2 are parametric. The parameter makes the discount function more flexible, allowing them to be used in many different decision situations. Therefore, in order to apply such functions in the information retrieval domain, the parameter of the functions needs to be optimized for the best performance. In this evaluation, different values for the parameters are used in order to discover the optimal value for each discount function. Table 5-3 shows the range of values for the parameter of each discount function to be used in this evaluation.

Table 5-3: The ranges of values for the parameter of the discount functions

Discount function	*Parameter*	*Range of values*
SAMUELSON	r	0.1 to 8.0
MAZUR	s	0.25 to 5.0
HARVEY	h	0.1 to 8.0

5.4 Experimental Results and Analysis

The role of the browsing recommendation model is to select one of the browsing paths extracted from the decision tree produced by the implicit RF system. Therefore, the decision tree produced by the implicit RF system from the experiments in Chapter 4 is used as the input to the browsing recommendation model in this evaluation (refer to Section 4.4.1 for the discussion on the experimental setup for the implicit RF system).

Similarly, there are three test collections used in this experiment, the TREC 1 ad-hoc retrieval collection (*TREC 1*), the TREC 7 ad-hoc retrieval collection (*TREC 7*) and the TREC 2003 .GOV collection with topic distillation tasks (*TREC .GOV*). Topics 51 to 100 are used for the TREC 1, topics 351 to 400 are used for the TREC 7 and the topic distillation set TD1 – TD50 is used for TREC .GOV.

5.4.1 Using the RSV values of the documents as the independent values

First, let us assume that the value of a document is independent and it is estimated based on the query submitted by the user. This assumption is generally accepted in IR research and it is used mainly to assign ranks to the retrieved documents. The value is often regarded as the *retrieval status value* (RSV) of the document. Such a value for a document is given by the value assigned by the Terrier IR system and it is computed based on the PL2 weighting model of the system with a default parameter setup, similar to the experiment in Chapter 4.

5.4.1.1 Tuning the parameter of the discount function for an optimal performance

The intertemporal choice browsing model is parametric due to the parameters of its value function and its discount function. The parameters are needed since the nature of the decision for the intertemporal choice problem is dynamic, such that it may change for different scenarios or for different times. For this experiment, the effect of different parameter values on the performance of the system is investigated and the best parameter's value for each discount function on each test collection is sought. Note that only the parameter for the discount function is tuned while the parameter for the value function is fixed. This is because the parameter for the value function will affect the value of the document perceived by the user and the user is not available in the simulation based experiment to verify such effects. In addition, the value of the document is given by Terrier IR system and the IR system is not being investigated in this experiment. Therefore, only the parameter value for the discount function is being tuned. In particular, the goal is to discover the best value of r for the SAMUELSON, h for the HARVEY and s for the MAZUR discount function.

Figure 5-3, Figure 5-4 and Figure 5-5 show the average MAP scores of the recommended browsing paths for the TREC 1, TREC 7 and TREC .GOV collections, respectively. Three different discount functions are used in the experiments, HARVEY, MAZUR and SAMUELSON. Different parameter values for the discount functions as stated in Table 5-3 are used in this experiment.

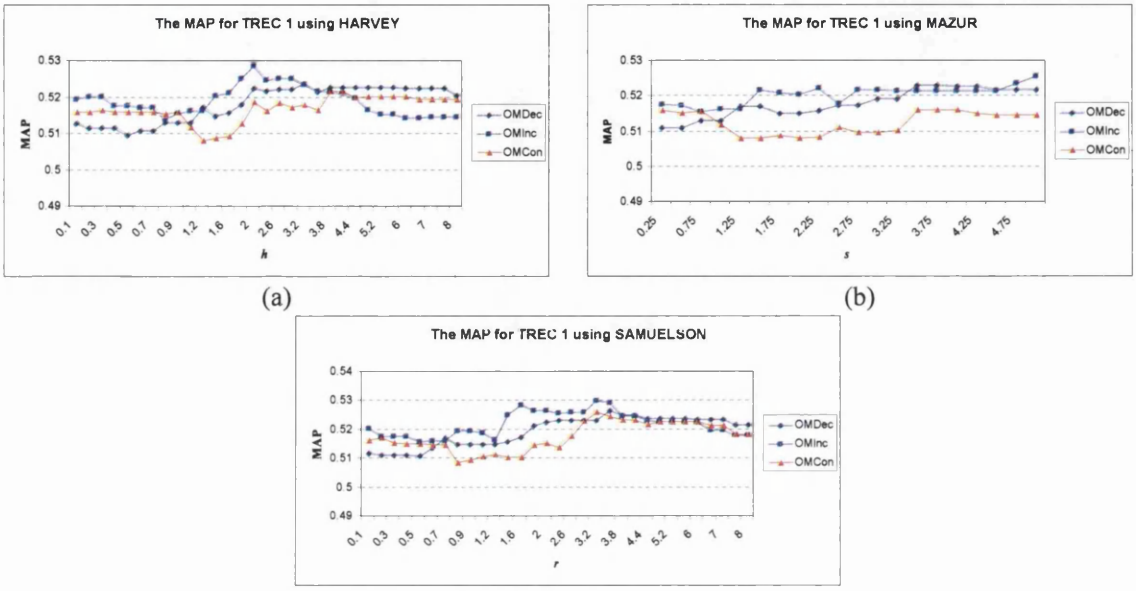


Figure 5-3: The average MAP scores of the recommended browsing path for TREC 1

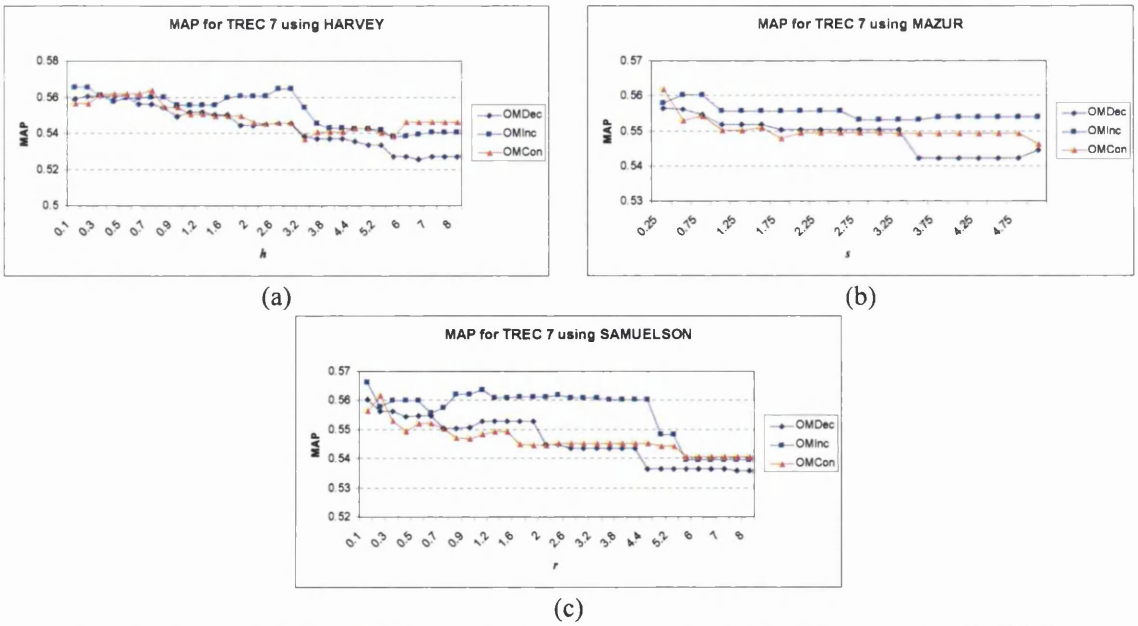
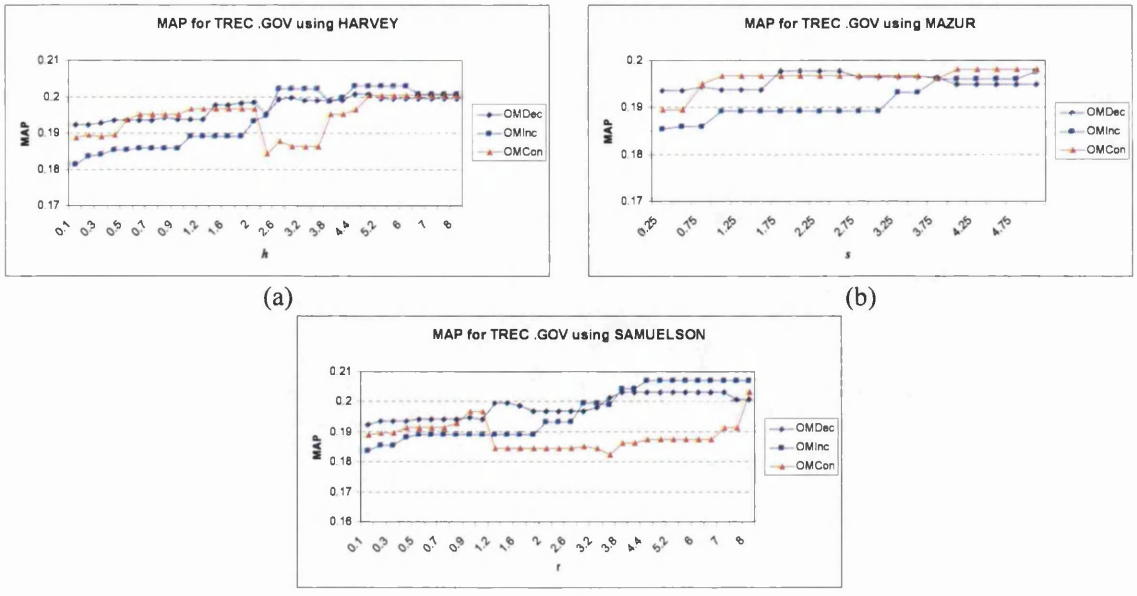


Figure 5-4: The average MAP scores of the recommended browsing path for TREC 7



(a) (b)

(c)

Figure 5-5: The average MAP scores of the recommended browsing path for TREC .GOV

Based on the graphs in Figure 5-3, Figure 5-4 and Figure 5-5, there is no obvious pattern of the impact of the MAP scores to the different parameter values for the discount functions. It is also difficult to determine the optimal setting for the discount functions. One interesting observation is that the MAP scores for the TREC 7 collection seem to gradually decrease as the parameter values increase, while the opposite patterns can be observed for the other collections. Based on this pattern, it may be wise to choose a higher parameter value of the discount functions for the TREC 1 and TREC .GOV collections, and, a lower parameter value for the TREC 7 collection.

Moreover, based on the graphs in Figure 5-3, Figure 5-4 and Figure 5-5, it is quite obvious that OMInc is the best implicit RF model for this evaluation. The browsing path recommended by the recommendation model for the OMInc model has the highest MAP score in almost test collections and discount functions. Therefore, OMInc is chosen as the implicit RF model for further experiments.

Since there is no obvious pattern on the effect of different parameter value on the performance of the system and the difference between the highest and the lowest MAP scores as a result of different parameter values is rather small, it is decided that the parameter's value of the discount function be fixed for further experiments. The best parameter value is fixed to its best performance for this evaluation. Therefore, the parameter of the discount functions is fixed to the value depicted in Table 5-4 that performs the best for OMInc.

Table 5-4: The best parameter value for OMInc

	Samuelson (*r*)	Harvey (*h*)	Mazur (*s*)
TREC 1	3.2	2.0	5.0
TREC 7	0.1	0.1	0.7
TREC .GOV	8.0	6.0	5.0

Next, the effectiveness of the recommendation model is compared to a few alternative strategies. The effectiveness of the recommendation model is measured by the browsing precision of the recommended browsing path. Note that, the recommendation engine selects one browsing path from all the possible browsing paths generated by the implicit RF system (OMInc) based on the intertemporal choice browsing model proposed in this thesis.

There are two alternative browsing strategies to be compared, the *random* browsing and the *simple* browsing. For the random browsing, the user chooses the document randomly each time, while for the simple browsing, the user will always choose the top document for browsing. The MAP score of the browsing paths as a result of the two strategies will be compared to the MAP score of the recommended browsing path.

Moreover, the performance of the implicit RF system is compared to the baseline that is based on the top ranking system. In Chapter 4, it has been shown that the implicit RF system may improve the effectiveness of the baseline by offering better browsing strategy for the user. However, due to the fact that there are a number of ways to browse the system, the system cannot guarantee to offer a better browsing strategy for the user. In this chapter, the role of the recommendation model is to select the best browsing path for the user. Therefore, it is interesting to compare the browsing precision of the recommended browsing path against the browsing precision of the baseline system.

5.4.2 The effectiveness of the recommendation model

In this section, the effectiveness of the recommendation model based on the intertemporal choice browsing model is evaluated. The effectiveness of the recommendation model is measured based on the mean average precision (MAP) of the recommended browsing path. Three discount functions to be used in the model, namely SAMUELSON, HARVEY and MAZUR, are also evaluated. The parameter of the discount function is set to its best performance discovered in the previous experiment.

5.4.2.1 The TREC 1 collection

In order to observe the relative performance of the model, the MAP score of the recommended browsing path is compared against the MAP score of the baseline, which is the score of the top-*20* retrieved documents. Based on this comparison, the real benefit of the implicit RF system in helping the user browsing can be observed. In addition, the performance of the recommendation model is also compared against two alternative browsing strategies, *random* browsing and *simple* browsing. A good model should recommend a browsing path that allows a user to have a better browsing experience compares to simple browsing and also random browsing.

Table 5-5: The MAP score of SAMUELSON for the TREC 1

Query quality	Samuelson $r = 3.2$	Baseline	Δ %	Average for OMInc	Δ %	Random	Δ %	Simple	Δ %
Excellent	0.921	0.906	1.59	0.914	0.74	0.911	1.02	0.934	-1.46
Good	0.660	0.667	-0.92	0.702	-5.86	0.706	-6.52	0.652	1.26
Moderate	0.351	0.345	1.73	0.346	1.28	0.341	2.82	0.317	10.75
Poor	0.080	0.094	-15.02	0.063	25.89	0.054	47.77	0.053	51.35
Overall	**0.530**	**0.528**	**0.27**	**0.530**	**0.02**	**0.518**	**2.19**	**0.518**	**2.19**

Table 5-6: The MAP score of HARVEY for the TREC 1

Query quality	Harvey $h = 2.0$	Baseline	Δ %	Average for OMInc	Δ %	Random	Δ %	Simple	Δ %
Excellent	0.921	0.906	1.58	0.914	0.73	0.911	1.01	0.934	-1.47
Good	0.647	0.667	-2.99	0.702	-7.83	0.706	-8.48	0.652	-0.86
Moderate	0.352	0.345	2.22	0.346	1.77	0.341	3.32	0.317	11.28
Poor	0.083	0.094	-11.66	0.063	30.87	0.054	53.62	0.053	57.34
Overall	**0.528**	**0.528**	**0.02**	**0.530**	**-0.23**	**0.518**	**1.93**	**0.518**	**1.93**

Table 5-7: The MAP score of MAZUR for the TREC 1

Query quality	Mazur $s = 5.0$	Baseline	Δ %	Average for OMInc	Δ %	Random	Δ %	Simple	Δ %
Excellent	0.926	0.906	2.15	0.914	1.29	0.911	1.58	0.934	-0.92
Good	0.658	0.667	-1.32	0.702	-6.24	0.706	-6.90	0.652	0.85
Moderate	0.347	0.345	0.69	0.346	0.25	0.341	1.77	0.317	9.62
Poor	0.061	0.094	-34.59	0.063	-3.11	0.054	13.74	0.053	16.49
Overall	**0.525**	**0.528**	**-0.54**	**0.530**	**-0.79**	**0.518**	**1.36**	**0.518**	**1.36**

Table 5-5, Table 5-6 and Table 5-7 show the MAP scores for the recommendation model that uses SAMUELSON, HARVEY and MAZUR as the discount function, respectively. All the percentage differences are computed against the MAP score of SAMUELSON, HARVEY or MAZUR, respectively. Based on the tables, the performance of the recommendation model is almost similar to the performance of the baseline. The performance of the model is also similar to the average performance of OMInc. It means the recommendation model chooses a browsing path that is as good as the average performance of the implicit RF system. These two findings are rather

disappointing since the benefit of the recommendation model as well as the implicit RF system can be neglected. In addition, the model recommends a browsing path that is better than random browsing path and simple browsing path. The MAP score of the recommended browsing path is about 2% better than the MAP score of random and simple browsing.

5.4.2.2 The TREC 7 collection

Next, the performance of the recommendation model is evaluated for the TREC 7 collection. Table 5-8, Table 5-9 and Table 5-10 show the MAP score of the recommendation model based on SAMUELSON, HARVEY and MAZUR, respectively, where the parameter value for the discount function is depicted accordingly and the value is set for optimal performance for the TREC 7 collection.

Table 5-8: The MAP score of SAMUELSON in the TREC 7

Query quality	Samuelson $r = 0.1$	Baseline	Δ %	Average for OMInc	Δ %	Random	Δ %	Simple	Δ %
Excellent	0.885	0.890	-0.53	0.913	-3.06	0.894	-0.96	0.899	-1.46
Good	0.763	0.574	32.98	0.750	1.78	0.674	13.22	0.641	19.17
Moderate	0.463	0.400	15.77	0.460	0.68	0.470	-1.55	0.441	4.86
Poor	0.095	0.084	13.46	0.058	63.50	0.088	7.45	0.086	10.80
Overall	**0.566**	**0.519***	**9.22**	**0.564**	**0.41**	**0.555**	**1.99**	**0.543**	**4.23**

Table 5-9: The MAP score of HARVEY in the TREC 7

Query quality	Harvey $h = 0.2$	Baseline	Δ %	Average for OMInc	Δ %	Random	Δ %	Simple	Δ %
Excellent	0.883	0.890	-0.76	0.913	-3.29	0.894	-1.19	0.899	-1.69
Good	0.763	0.574	32.98	0.750	1.78	0.674	13.22	0.641	19.17
Moderate	0.463	0.400	15.77	0.460	0.68	0.470	-1.55	0.441	4.86
Poor	0.095	0.084	13.46	0.058	63.50	0.088	7.45	0.086	10.80
Overall	**0.566**	**0.519***	**9.08**	**0.564**	**0.28**	**0.555**	**1.86**	**0.543**	**4.10**

Table 5-10: The MAP score of MAZUR in the TREC 7

Query quality	Mazur $s = 0.75$	Baseline	Δ %	Average for OMInc	Δ %	Random	Δ %	Simple	Δ %
Excellent	0.888	0.890	-0.19	0.913	-2.73	0.894	-0.62	0.899	-1.13
Good	0.729	0.574	26.99	0.750	-2.81	0.674	8.11	0.641	13.80
Moderate	0.465	0.400	16.24	0.460	1.08	0.470	-1.16	0.441	5.28
Poor	0.086	0.084	2.77	0.058	48.08	0.088	-2.68	0.086	0.36
Overall	**0.560**	**0.519***	**8.03**	**0.564**	**-0.68**	**0.555**	**0.88**	**0.543**	**3.10**

Based on the above tables, the performance of the recommendation model is significantly better than the baseline. The percentage difference of 9.22% for the SAMUELSON (p=0.0255, paired t-test), 9.08% for the HARVEY (p= 0.0274, paired t-test) and 8.03% for the MAZUR (p= 0.0456, paired t-test) are significant. Similar to the results for the TREC 1 collection, the model recommends a browsing path that is as

good as the average performance of the implicit RF system (OMInc). These two results mean that the recommendation model is useful since it recommends a better browsing path to the user (as compared to the baseline) but it fails to improve the average performance of the implicit RF system itself.

Similar to the TREC 1 collection, the performance of the recommendation model is slightly better than random and simple browsing. The percentage difference is about 1% to 2% for random browsing and is about 3% to 4% for simple browsing. It means that following the recommended browsing path is still better than random browsing or simple browsing.

5.4.2.3 The TREC .GOV collection

Table 5-11, Table 5-12 and Table 5-13 show the MAP scores for the recommendation model evaluated on the TREC .GOV collection. Based on the tables, the performance of the recommendation model is slightly better than the baseline. The percentage difference of 6.76% for SAMUELSON, 4.54% for HARVEY and 1.9% for MAZUR are regrettably insignificant.

Table 5-11: The MAP score of SAMUELSON for the TREC .GOV

Query quality	Samuelson $r = 8.0$	Baseline	Δ %	Average for OMInc	Δ %	Random	Δ %	Simple	Δ %
Excellent	0.412	0.934	-55.88	0.392	5.19	0.335	22.86	0.354	16.34
Good	0.000	0.500	-100	0.000	0.0	0.000	0.0	0.000	0.0
Moderate	0.259	0.355	-26.89	0.204	26.92	0.208	24.91	0.236	9.73
Poor	0.173	0.047	267.68	0.157	9.98	0.135	28.32	0.165	4.42
Overall	**0.207**	**0.194**	**6.76**	**0.185**	**11.96**	**0.164**	**26.51**	**0.192**	**7.66**

Table 5-12: The MAP score of HARVEY for the TREC .GOV

Query quality	Harvey $h = 6.0$	Baseline	Δ %	Average for OMInc	Δ %	Random	Δ %	Simple	Δ %
Excellent	0.412	0.934	-55.88	0.392	5.19	0.335	22.86	0.354	16.34
Good	0.000	0.500	-100	0.000	0.0	0.000	0.0	0.000	0.0
Moderate	0.259	0.355	-26.89	0.204	26.92	0.208	24.91	0.236	9.73
Poor	0.167	0.047	254.91	0.157	6.16	0.135	23.86	0.165	0.80
Overall	**0.203**	**0.194**	**4.54**	**0.185**	**9.63**	**0.164**	**23.87**	**0.192**	**5.42**

Table 5-13: The MAP score of MAZUR for the TREC .GOV

Query quality	Mazur $s = 5.0$	Baseline	Δ %	Average for OMInc	Δ %	Random	Δ %	Simple	Δ %
Excellent	0.382	0.934	-59.13	0.392	-2.56	0.335	13.81	0.354	7.77
Good	0.000	0.500	-100	0.000	0.0	0.000	0.0	0.000	0.0
Moderate	0.259	0.355	-26.89	0.204	26.92	0.208	24.91	0.236	9.73
Poor	0.164	0.047	248.75	0.157	4.32	0.135	21.71	0.165	-0.95
Overall	**0.198**	**0.194**	**1.90**	**0.185**	**6.86**	**0.164**	**20.74**	**0.192**	**2.75**

In the case of the poor queries, the recommendation model tremendously improves the performance of the baseline. The improvement of 268% for the SAMUELSON (p=

0.0194, paired *t*-test), 255% for the HARVEY (p= 0.0267, paired *t*-test) and 249% for the MAZUR (p= 0.0288, paired *t*-test) are significant.

In addition, the recommended browsing path is much better than the average performance of the implicit RF system with 7% to 12% improvement. The performance of the recommendation model is also 3% to 8% better than the performance of simple browsing, while it is around 21% to 27% better than random browsing.

By looking at the MAP scores for the discount functions in the three test collections, it is obvious that the performance of the three discount functions is quite similar. SAMUELSON is probably a little bit better than HARVEY and MAZUR, but the difference is insignificant. Based on this result, the hypothesis that the hyperbolic discount functions (HARVEY and MAZUR) are better than the exponential discount function can be rejected in this context of this evaluation (refer to Section 3.3.4.3 for the discussion on the shape of the discount function).

5.4.2.4 The expected search length (ESL) of the recommendation model

In this section, the performance of the recommendation model is analysed based on the expected search length (ESL). The parameter of the discount function is set to the same value as the previous experiments (its optimal value). The queries are divided into two categories. The first category consists of those queries with at least one relevant document retrieved at the top-*20* by the baseline system. The ESL score, which is the expected no of non-relevant document assessed before finding the first relevant document, can be computed for these queries. The second category consists of those queries with no relevant document retrieved at the top-*20*.

Table 5-14: The ESL score for the queries with at least one retrieved relevant document

	Baseline	Simple	Samuelson	Harvey	Mazur
TREC 1	2.48	4.13	3.47	3.47	3.73
TREC 7	2.42	3.14	2.99	2.99	3.14
TREC .GOV	5.02	12.37	11.12	11.69	10.97

The ESL score for the queries in the first category is depicted in Table 5-14. Based on the table, it is obvious that the recommendation model fails to improve the performance of the baseline. The ESL score for the baseline is lower than the ESL score for the

recommendation model. Meanwhile, the performance of the recommendation model is slightly better than the performance of simple browsing.

The ESL of the queries in the second category is reported in Table 5-15. In this case, the recommendation model helps the user by offering a better ESL score as compared to the baseline. It means that the implicit RF system with recommendation is able to retrieved relevant documents when the baseline system fails to retrieve them.

Table 5-15: The ESL score for the queries with no retrieved relevant document

	Baseline	Simple	Samuelson	Harvey	Mazur
TREC 1	20+	20+	19.38	17.38	17.38
TREC 7	20+	12.75	12.75	12.75	12.75
TREC .GOV	20+	12.30	11.91	11.91	12.68

5.4.3 The effectiveness of the DecreasingWeight and the EqualWeight techniques

In Section 5.2.2, three techniques to construct a sequence of values from a browsing path are described, which are *SingleValue*, *EqualWeight* and *DecreasingWeight*. The aim of the techniques is to construct a sequence of values from the documents in the browsing path. As a recap, *SingleValue* takes only the value of the chosen documents in a browsing session as the sequence of values for the path, while *EqualWeight* and *DecreasingWeight* take into consideration the values of all displayed documents.

In the previous experiments in this chapter, *SingleValue* technique has been used to construct a sequence of values for the browsing paths. In this section, the effectiveness of the *EqualWeight* and *DecreasingWeight* techniques will be evaluated and its performance will be compared to the performance of the *SingleValue* technique. OMInc is used as the implicit RF system in this experiment. The recommendation is computed based on these three techniques with three different discount functions (HARVEY, MAZUR and SAMUELSON) and with different parameter values for the discount function. The average MAP scores for different parameter values of the discount function is computed and it is used for the comparison. The parameter values chosen are within the ranges of values depicted in Table 5-3.

Table 5-16, Table 5-17 and Table 5-18 show the average MAP scores of the recommendation model using the three techniques with HARVEY, MAZUR and

SAMUELSON as the discount functions, respectively. In general, the *DecreasingWeight* and the *EqualWeight* techniques perform better than the *SingleValue* technique for the TREC 1 and the TREC 7 collection, while for the TREC .GOV collection, the performance is otherwise. The difference in the average MAP scores are all significant ($p \approx 0.0001$, paired t-test), indicated by (*) sign.

Table 5-16: The average MAP for HARVEY

	SingleValue	DecreasingWeight	Δ %	EqualWeight	Δ %
TREC 1	0.519	0.528*	1.81	0.526*	1.45
TREC 7	0.553	0.566*	2.48	0.567*	2.63
TREC .GOV	0.194	0.185*	-4.88	0.184*	-5.24

Table 5-17: The average MAP for MAZUR

	SingleValue	DecreasingWeight	Δ %	EqualWeight	Δ %
TREC 1	0.520	0.526*	1.17	0.529*	1.76
TREC 7	0.555	0.570*	2.70	0.566*	1.94
TREC .GOV	0.195	0.183*	-6.48	0.185*	-5.49

Table 5-18: The average MAP for SAMUELSON

	SingleValue	DecreasingWeight	Δ %	EqualWeight	Δ %
TREC 1	0.522	0.528*	1.19	0.525*	0.70
TREC 7	0.556	0.566*	1.72	0.567*	2.04
TREC .GOV	0.196	0.185*	-5.65	0.183*	-6.61

The *DecreasingWeight* and the *EqualWeight* techniques should be better than the *SingleValue* technique since the value of all displayed documents is taken into consideration during the construction of the browsing paths. However, the statement is not true for the TREC .GOV collection. Moreover, the difference between the *DecreasingWeight* technique and the *EqualWeight* technique is insignificant.

5.4.4 Using the QRel values as the independent values for the documents

One of the main problems discussed concerning the intertemporal choice browsing model is the value estimation of the documents. As commonly understood, the value of a document should be estimated based on the information need or the query of the user, which is also known in this thesis as the independent value for the document. So far in this chapter, the retrieval status value (RSV) computed by the IR system is used as the value of a document. Therefore, the performance of the intertemporal choice browsing model depends on the quality of the RSV value given by the IR system.

Meanwhile, in the test collections investigated in this chapter, each document is assigned a QRel value for a given topic. In particular, a document is assigned with a value of 1 if it is relevant to that topic and a value of 0 if it is not relevant to that topic. In addition, such relevance judgement is done by the people that suggested the topic. Assuming that those topics or queries represent the real information need of the people and the relevance judgements on the documents are made by those people, the relevance value assigned to the documents (1 or 0) can be treated as the real value of the documents.

In order to investigate the actual capability of the intertemporal choice browsing model as the browsing recommendation model, the QRel values are used for the documents. In this section, some of the previous experiments conducted in this chapter are repeated with the QRel values assigned to the documents. In the next section, the effectiveness of using the QRel values for the documents is investigated and compared with the RSV values. The effectiveness of using the QRel values is also compared against the effectiveness of the baseline system. After that, the effect of using the QRel values is investigated in the context of the *SingleValue*, the *EqualWeight* and the *DecreasingWeight* techniques. The effectiveness of the system is measured by using both the MAP score and the ESL score.

First, the effect of using QRel values as the values for the documents is investigated for the three discount functions. Table 5-19 shows the average MAP scores for the recommended browsing paths, computed for different parameter values of the discount functions. The average MAP score indicates the overall performance of the recommendation model for different values assigned to the parameter. The RSV columns show the average MAP scores when the retrieval status values are used as the value of the document, while the QRel columns show the average MAP scores when the QRel values, taken from the test collection, are used as documents' values.

Table 5-19: The average MAP scores when QRel and RSV values are used

	HARVEY			MAZUR			SAMUELSON		
	RSV	QRel	Δ %	RSV	QRel	Δ %	RSV	QRel	Δ %
TREC 1	0.519	0.564	8.75	0.520	0.564	8.46	0.522	0.564	8.17
TREC 7	0.553	0.564	2.08	0.555	0.564	1.50	0.556	0.565	1.60
TREC .GOV	0.194	0.222	14.26	0.195	0.222	13.39	0.196	0.222	13.32

Based on Table 5-19, it is obvious that the QRel values are better than the RSV values to represent the value of documents. The difference is about 8% for the TREC 1 collection, 2% for the TREC 7 collection and 13%-14% for the TREC .GOV collection. The difference is considered marginal for the TREC 1 and TREC 7 collection. Even though, such a pattern should be expected since the QRel represents the true value of relevance for the documents to a certain extends. The marginal difference between the use of RSV and QRel describes that the retrieval status value assigned by the search engine to the documents is quite good.

Table 5-20: The best MAP score of using QRel

	Upper bound	HARVEY	Δ %	MAZUR	Δ %	SAMUELSON	Δ %
TREC 1	0.633	0.564	-10.79	0.564	-10.79	0.564	-10.79
TREC 7	0.656	0.565	-13.82	0.565	-13.82	0.567	-13.54
TREC .GOV	0.251	0.223	-11.18	0.223	-11.18	0.223	-11.18

Second, the best MAP score of the recommendation model achieved by using the QRel values is compared to the maximum MAP score (upper bound) achievable by the user browsing through the implicit RF system (OMInc) and it is shown in Table 5-20. Based on this result, the actual capability of the browsing recommendation model in suggesting the best browsing path can be observed. Based on the table, the recommendation model fails to suggest the best browsing path for the user even with the use of QRel. The difference is about 10% for the TREC 1 collection, 14% for the TREC 7 collection and 11% for the TREC .GOV collection.

5.5 Discussion

The aim of this chapter is to evaluate the performance of the proposed recommendation model for the implicit RF system discussed in Chapter 4. In the previous chapter, it has been found that the implicit RF system has a good potential to improve the performance of the baseline system. However, the system needs a recommendation model to suggest a good browsing path to the user. Therefore, the intertemporal choice browsing model discussed in Chapter 3 is used as the recommendation model for the implicit RF system.

5.5.1 The effectiveness of the recommendation model

In Section 5.4.1.1, the performance of the recommendation model is evaluated for different values of the discount functions. The purpose of the experiment is to discover the relationship between the parameters' value and the performance of the model and also to determine the optimal value for those parameters. However, there is no obvious

pattern can be discovered from the result of the experiment. Therefore, it becomes another challenge to find the optimal setting for the recommendation model. Moreover, the optimal setting of the parameter may depend on the test collection as shown in Figure 5-3, Figure 5-4 and Figure 5-5. The problem of finding the optimal setting of the model will not be further discussed in this thesis. For further experiments, the value of the parameter is set to its best performance based on the result of this experiment.

By using the optimal setting for the recommendation model, the performance of the model is compared to the performance of *simple* browsing and *random* browsing. In the simple browsing, the user will always choose the top displayed document while in the random browsing, the document is randomly chosen. The experiment shows that in this evaluation setting the performance of the recommendation model is slightly better than the performance of random or simple browsing. It means that it is still reasonable to follow the recommendation of the model. However, the recommendation given is not necessarily better than the baseline. A significant improvement can only be observed in one of the test collections. Moreover, the quality of the recommended browsing path, measured by the MAP score, is as good as the average performance of the implicit RF system. Based on these findings, the recommendation model proposed in this thesis is still a reasonable solution to be incorporated into the implicit RF system. However, further improvement of the recommendation model is required to improve the overall performance of the implicit RF system discussed in Chapter 4.

However, the recommendation model does not improve the performance of the baseline in term of the expected search length (ESL) score to find the first relevant document. Based on the results, the user will have to assess more non-relevant documents before reaching to a relevant document. Meanwhile, the recommendation model can help the user finding relevant documents for those queries with no relevant documents retrieved at the top-*20* by the baseline.

5.5.2 The effect of modelling uncertainty of the outcomes

In Section 3.3.3.4 of Chapter 3, the need to model uncertainty for the value of the outcomes in the intertemporal choice browsing model has been discussed. It is based on the assumption that in the case of IR, the value of an outcome at a given time cannot be estimated with certainty. In Section 3.3.3.4, two scenarios where this condition applies have been discussed, which are the uncertainty in the context of the query of which the

value of a document is estimated and the situation where there are more than one document received by the user at a given time. A detailed discussion of this aspect is presented in Section 3.3.3.4.

In this evaluation, the effect of modelling uncertainty of the outcomes for the second scenario has been investigated, where there are *four* documents displayed to the user each time. In Section 5.2.2, two strategies to estimate the probabilistic weight of each document in a given sequence of browsing were presented, an equal weight and a decreasing weight with respect to the ranks. Based on the experimental results in Section 5.4.3, it can be observed that the performance of the *EqualWeight* and the *DecreasingWeight* techniques may not necessarily better than the performance of the *SingleValue* technique and also it depends on the test collection.

5.5.3 The effectiveness of the discount functions

In many empirical studies of the intertemporal choice model, the hyperbolic discount function seems to be better than the exponential discount function in modelling the behaviour of an individual (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Lazaro et al., 2002). Due to this reason, the effectiveness of one exponential discount function and two hyperbolic discount functions to be used for the intertemporal choice browsing model is investigated. However, based on the experimental results, the effectiveness of using the three discount functions is similar, which indicates that there is no difference in the effectiveness of the intertemporal choice browsing model for different shape of the discount function. Therefore, any discount function can be used for the intertemporal choice model in the context of this application. Further investigation on the discount function should be focused on finding the optimal setting of the discount function.

5.5.4 Using QRel values for perfect recommendation

The main reason to conduct an evaluation of the intertemporal choice browsing model based on using the QRel values is to investigate the performance of the model in the context of perfect value estimation for the documents. The aims is to eliminate the dependency of the browsing recommendation model on the retrieval status value (RSV) produced by an IR system.

In such a perfect setting, the performance of the recommendation model is not necessarily big. In one of the test collections, only a marginal improvement of 2% is observed. In that case, the retrieval status value (RSV) is already a good value to be used for the documents. In other test collections, the difference in the performance is about 8% to 14%. However, in that setting, the recommendation model is unable to produce the best browsing path for the user. The performance of the recommendation model in such a perfect setting is still about 10% to 14% less than its optimal performance.

The use of QRel as the value for the documents should be considered as the perfect setting since it is also used in the computation of the evaluation measure in this experiment. Moreover, the parameter value of the discount function is carefully tuned for the best performance. However, the intertemporal choice browsing model still fails to find the best browsing path. Based on these findings, it is safe to assume that the intertemporal choice browsing model may not be suitable for this application in the context of this evaluation. A user based study should be used to further investigate the potential capability of this model.

5.6 Chapter Summary

In this chapter, the effectiveness of the intertemporal choice browsing model to predict a good browsing path for the user of the implicit RF system has been investigated. Some of the implementation issues regarding the use of the intertemporal choice browsing model as the browsing recommendation model for the implicit RF system has also been discussed. Based on the evaluation, the intertemporal choice browsing model proposed appears to be effective as a browsing recommendation model for the implicit RF system investigated. However, its performance is still far from perfect and further investigation should be focused on finding the optimal setting for the model. In the next chapter, the effectiveness of the intertemporal choice browsing model is investigated for another application in IR which is the post-retrieval browsing in the context of the subtopic relevance retrieval.

6 Re-ranking of the Top-*n* Documents for Post Retrieval Browsing

6.1 Introduction

A common way to present the results of a search is through a ranked list. The retrieved documents are ranked in decreasing order of their relevance value or their retrieval status value (RSV). Such ranking technique follows the principle of the Probability Ranking Principle (PRP) (Robertson, 1977). The top-*n* documents that are shown first on the search result page are quite important since the user is likely to browse from these documents to find relevant information. The browsing of the top-*n* retrieved documents is described as *post-retrieval browsing* in this chapter.

Post-retrieval browsing occurs when a user is assessing the retrieved documents sequentially to find information that is relevant to his/her information need. The sequence in which the documents are assessed is important in browsing. Intuitively, a good browsing sequence or browsing strategy should lead the user to the relevant information and it should require fewer documents to be assessed or less effort from the user.

A common understanding by the user regarding the ranked list is that a higher rank is always better. Therefore, a user is likely to browse the list according to the ranks, starting from the document at the top of the list. It is considered a reasonable browsing strategy, since the rank assigned to a document depends on its usefulness to the user as predicted by the search engine. The ranking also acts as an indirect recommendation to

the user. Therefore, the quality of the ranking produced determines the effectiveness of the browsing strategy.

It has been argued that the PRP is the best ranking technique since it optimises the expectation of finding the relevant information or relevant documents (Robertson, 1977). The technique is also examined in a utility theoretic framework which suggested that the PRP will be optimal provided the following three conditions are not violated by the information retrieval (IR) system; (1) the predicted probability of relevance for a document is well-calibrated, (2) it is reported with certainty and (3) the documents are assessed independently (Gordon and Lenk, 1991).

Due to the fact that during post-retrieval browsing, the documents are assessed in sequence, the condition of independent assessment is violated. In this scenario, the value of the next document to be assessed by a user depends on the documents he/she has already read. In this chapter, this concept is described as the *dependent relevance assumption.* In (Gordon and Lenk, 1992), the authors suggested that any violation of the condition will make the ranking become sub-optimal. Robertson has also anticipated this problem in (Robertson, 1977).

A browsing strategy that follows the ranking of the top-*n* documents will not be a good strategy due to the fact that the ranking itself is sub-optimal. However, the problem may not be obvious for an IR system that is based on the *topical relevance retrieval* application. The problem becomes more obvious in the context of the *subtopic relevance retrieval* application (Zhai et al., 2003).

Topical relevance retrieval is the traditional retrieval problem that deals with the problem of finding a set of documents that are relevant to a particular topic. Such documents are assessed based on their relevance to a given topic of interest. A document is either relevant or irrelevant to the topic. On the other hand, subtopic relevance retrieval deals with the problem beyond the topical relevance scenario (Zhai et al., 2003). In this case, a given topic consists of a set of subtopics. Each document is judged based on its cumulative relevance to each subtopic. A document can be relevant to none and up to all subtopics of a given topic. This retrieval problem is concerned more with how many subtopics of a given topic are covered by the set of retrieved documents.

In browsing, the potential of finding relevant documents is not the only important issue. In addition, browsing is also concerned about the effort spent on finding relevant documents. A shorter browsing length (or browsing time) to find the intended information is always better than a longer one. Usually, there is a trade-off between finding a *more relevant* document and the time or the cost of the longer browsing length. The question is whether a user should continue to look for a more relevant document or whether he/she should stop and be satisfied with a less relevant document.

The notion of more or less relevant is usually vague and it is difficult to model. However, without this notion, it is very difficult to justify the trade-off of browsing. Therefore, the problem of finding an optimal ranking or browsing strategy will be discussed in the context of the subtopic relevance retrieval application. For this application, it is assumed that for a given topic, a document is more relevant to the user if it contains more relevant subtopics.

The aim of this chapter is to improve the ranking of the top-*n* documents for the purpose of providing an optimal browsing strategy, which facilitates a better browsing experience for the user. The top-*n* documents are chosen assuming that these documents are likely to be assessed by the user. Also, a user is unlikely to browse through the entire set of documents in the collection. In addition, a smaller set of documents to be re-ranked will ensure an efficient implementation of the system. Therefore, re-ranking of the top-*n* documents for an optimal post retrieval browsing is reasonable.

The subtopic relevance retrieval application offers another challenge to the conventional ranking method. A document can no longer be classified as either relevant or non-relevant to a given topic. A document can be relevant to any set of subtopics. Furthermore, the system should take into consideration the novelty aspect of the ranking in the context of sequential browsing.

The Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) ranking method has been suggested as a solution for the subtopic relevance retrieval problem (Zhai et al., 2003). The technique combines both the relevance score and the novelty score into a single score to be used in the ranking. A variation of the MMR technique is

proposed by Zhai *et al.* for their Risk Minimization Framework for IR to deal with the same problem (Zhai et al., 2003). In this chapter, the effect of the technique on the post-retrieval browsing problem will be observed in the context of the subtopic relevance retrieval application.

In addition, the application of the intertemporal choice browsing model to the post-retrieval browsing problem will also be discussed. It is an attempt to optimise the MMR ranking method by using an exhaustive approach to the selection of the ranking as opposed to the greedy approach adopted by the MMR ranking method. The model also captures the trade-off between finding a more relevant document and the effort spent for browsing.

The chapter is structured as follows. In Section 6.2, the use of the MMR ranking method to produce an effective ranking for the subtopic relevance retrieval application is discussed. A number of existing novelty models to be used for computing the novelty score in the MMR ranking method will be presented. Then, the application of the intertemporal choice browsing model as an optimisation strategy for the problem of subtopic relevance retrieval will be discussed. In Section 6.3, the evaluation framework to be used in this chapter is described. In Section 6.4, the results of the experiments are presented. The evaluation results are discussed in Section 6.5. Finally, a brief summary for the chapter is provided in Section 6.6.

6.2 Subtopic relevance retrieval model

In the subtopic relevance retrieval application, *novelty* plays an important role. Novelty refers to the new information contained in a document measured with respect to the documents already viewed by a user. An effective subtopic relevance retrieval application should promote novelty in its ranking. This new feature is added to the conventional ranking method for the subtopic relevance retrieval application.

In (Zhai et al., 2003), the authors indicated that the Maximal Marginal Relevance (MMR) ranking method may be suitable for the subtopic relevance retrieval application. The MMR ranking method, proposed by Carbonell and Goldstein, takes into account the relevance score of a document against the topic and the novelty score of that document against the documents at higher ranks.

6.2.1 Maximal Marginal Relevance (MMR)

The score of the documents in the MMR ranking method depends on two components, the relevance score and the novelty score. The relevance score is measured based on the topic or the query statement submitted by the user. It is also called the retrieval status value (RSV) of a document. In the conventional ranking method, the relevance score or the RSV is used to produce the ranking. The novelty score of a document measures the new information contained in the document with respect to the other documents. Usually it does not depend on the topic or the query of the user. The MMR ranking method combines both scores as the final score to rank the documents. In (Carbonell and Goldstein, 1998), the relevance score and the novelty score are linearly combined.

The MMR ranking method is a greedy approach. Initially, the *ranklist* is empty and the retrieval engine assigns a relevance score to each retrieved document. At this point, the novelty score for each document is not defined since the *ranklist* is still empty. The novelty score for a document is computed based on the documents already in the *ranklist*. Therefore, the document with the highest relevance score is chosen for the first rank. The novelty score is then computed for all remaining documents in the retrieved set with respect to the documents already in the ranking. For each document, the MMR score is computed by linearly combining the relevance score and the novelty score. The document with the highest MMR score or combined score is chosen for the second rank. The process continues until all documents in the retrieved have been ranked.

6.2.2 The novelty models

In order to apply the MMR ranking method, the system needs a model to compute the novelty score for a document. In the literature, there are many models designed to measure the novelty of an information object such as documents and sentences (Zhang et al., 2002, Allan et al., 2003, Gabrilovich et al., 2004). The novelty models studied in this chapter can be divided into two categories, *distance-based* and *set-based* novelty models.

Let $novelty(d_i)$ be the novelty score for document d_i. Since the novelty score of a document is computed with respect to the other documents, the score can be computed based on the following equation:

$$novelty(d_i) = novel(d_i \mid d_1,...,d_{i-1}) \quad (6\text{-}1)$$

The function $novel(d_i \mid d_1,...,d_{i-1})$ computes the novelty score of document d_i with respect to the set of previous documents $d_1,...,d_{i-1}$. A number of methods have been proposed to model $novel(d_i \mid d_1,...,d_{i-1})$.

6.2.2.1 Distance-based novelty models

The basic idea of distance-based novelty models is to measure the novelty of a document based on its distance from the set of previous documents. The distance between d_i and each document in the set $d_1,...,d_{i-1}$ is computed based on Equation (6-1). The novelty score of the document can be based on the *average* distance or the *minimum* distance of that document to all documents in the set.

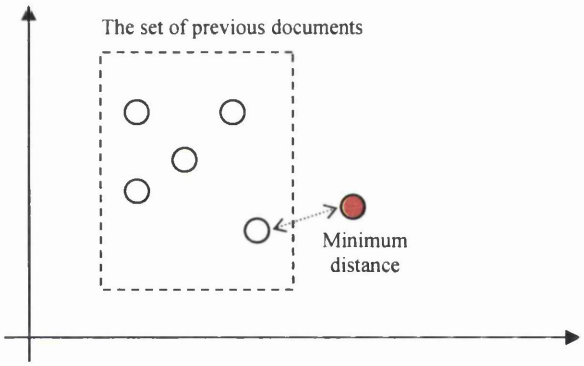


Figure 6-1: The minimum distance computed based on the set of previous documents

Figure 6-1 illustrates the concept of minimum distance as a novelty score for a document. The novelty score of the document (illustrated as a read circle) is computed as its minimum distance to the other previous documents. In this case, the novelty score is represented by the difference among the documents and the maximum difference of the documents can be represented by the minimum distance of the new document with the previous documents. Alternatively, the average distance of the new document to the previous documents can also be used as the novelty measure. The average distance is the average difference between the new document and the previous documents. In this case, the minimum distance should be a better novelty measure as compared to the average distance.

Table 6-1: The list of distance-based novelty models

ID	*Novelty Model*	*Description*
DN1	COSINE_AVG	Average cosine distance
DN2	COSINE_MIN	Minimum cosine distance
DN3	KL_AVG	Average Kullback-Leibler similarity
DN4	KL_MAX	Maximum Kullback-Leibler similarity

Based on the idea, a few distance-based novelty models have been proposed in the literature. In this chapter, the models that are based on the *cosine* distance measure and the *Kullback-Leibler* similarity measure will be investigated (Zhang et al., 2002).

Table 6-1 shows four distance-based novelty models that are being investigated in this chapter. The DN1 and the DN2 models are based on the cosine distance measure and the DN3 and the DN4 models are based on the *Kullback-Leibler* similarity measure.

6.2.2.2 Set-based novelty models

Set-based novelty models treat a document as a bag of words or a set of words. Accordingly, the novelty measure is derived from the difference between the sets of words. For instance, the novelty of a document d_i with respect to the set of documents $d_1,...,d_{i-1}$ is measured based on the difference between the set of words in d_i and the set of words in $d_1,...,d_{i-1}$.

One of the techniques is to measure the new words contained in the document d_i as compared to the words in $d_1,...,d_{i-1}$ (Allan et al., 2003). The following equation is one of the most popular novelty models and it has been shown to perform the best. The novelty score for document d_i is measured by the number of unique words in the document that do not appear in document $d_1,...,d_{i-1}$.

$$\text{SN1: } newWord(d_i \mid d_1,...,d_{i-1}) = \left| d_i \cap \overline{\bigcup_{j=1}^{i-1} d_j} \right| \tag{6-2}$$

In (Gabrilovich et al., 2004), the authors proposed a variation of the new word novelty model. They suggested that the number of new words in the documents is normalized to the total number of words in the document such that:

$$\text{SN2: } normalizedNewWord(d_i \mid d_1,...,d_{i-1}) = \frac{newWord(d_i \mid d_1,...,d_{i-1})}{|d_i|} \quad (6\text{-}3)$$

The set difference model (SN3) is another version of the simple new word model, where each word is assigned a weight and only those words with weight exceeding a certain threshold will be counted in the calculation (Allan et al., 2003). Based on Equation (6-4), the novelty score of a document is the number of new words in the document that do not appear in the other previous documents. In the case of the set difference model, a word w_m is considered belong to document d, $w_m \in d$, if and only if $count(w_m, d) > k$, where:

$$count(w_m, d) = (\alpha_1 \times tf_{w_m,d}) + (\alpha_2 \times df_{w_m}) + (\alpha_3 \times rdf_{w_m}) \quad (6\text{-}4)$$

$tf_{w_m,d}$ is the frequency of word w_m in document d, df_{w_m} is the number of documents in the top-*n* that contain w_m and rdf_{w_m} is the number of relevant documents that contain word w_m. In the experiment, α_1 is set to 0.8, α_2 is set to 0.2 and k is set to 2. Since the relevant documents are unknown, the parameter α_3 is set to 0.0 to eliminate the effect of rdf_{w_m}. These setting is also used by the authors in (Zhang et al., 2002).

The overlap is the ratio of the number of overlapping words between two documents and the length of the document (Zhang et al., 2004). The overlap novelty measure is computed as follows:

$$overlapScore(d_i, d_j) = \frac{|d_i \cap d_j|}{|d_i|} \quad (6\text{-}5)$$

$$\text{SN4: } overlap(d_i \mid d_1,...,d_{i-1}) = 1.0 - \max_{j=1}^{i-1} overlapScore(d_i, d_j) \quad (6\text{-}6)$$

Table 6-2 shows the list of the set-based novelty models to be investigated in this chapter. The models will be referred to as SN1, SN2, SN3 and SN4 in this chapter.

Table 6-2: The list of the set-based novelty models

ID	*Novelty Model*	*Description*
SN1	NEW_WORD	New word count
SN2	NORMALISED_NEW_WORD	New word count normalised by the length
SN3	SET_DIFFERENCE	Weighted word count difference
SN3	OVERLAP	One minus the overlap ratio

6.2.3 The intertemporal choice browsing model and ranking optimization

The strength of the MMR ranking method is that it incorporates relevance and novelty information in the ranking of documents. By assuming that the user satisfaction while browsing through the top-*n* retrieved documents depends on the effectiveness of the system in reducing redundancy and promoting novelty in the ranking, we turn our attention to the MMR ranking method for providing a solution to the problem. However, due to fact that it is a greedy approach (refer to Section 6.2.1), the method may not be optimal since it does not take into consideration all possible rankings of the documents. A more exhaustive approach can be adopted as a better solution, where the best ranking is be chosen from all possible rankings of the documents. In order to apply such approach, each possible ranking candidate should be assigned with a certain value or score. Such score should be based on the importance of the ranking or the sequence of documents to the user.

The intertemporal choice browsing model is capable of assigning a value or a score to a sequence of documents or a ranking based on its importance to the user. The model is based on the trade-off between the cost of browsing and the relevance of the documents that the user is about to find through browsing. The model is discussed in detail in Chapter 3.

Let $\{d_1,t_1;...;d_n,t_n\}$ be a sequence of documents where d_i is the document at time t_i and u_i is the value for document d_i. Recall from Chapter 3, the value for the sequence of documents is given by the following equation:

$$U_k = U(u_1,t_1;...;u_n,t_n) = \sum_{i=1}^{n} v(u_i)\phi(t_i) \qquad (6\text{-}7)$$

$v(u_i)$ is the value function that transforms the objective value for the document, u_i, into a subjective value and $\phi(t_i)$ is the discount function applied to t_i. The value function is

a monotonic increasing concave function for a positive u_i and a monotonic decreasing convex function for a negative u_i. The discount function is an exponential or a hyperbolic decreasing function for time t_i. The discount functions to be investigated are depicted in Table 5-2 as described in (Cairns and van der Pol, 2000, Lazaro et al., 2002), the same functions used in the experiments in Chapter 5.

6.3 Evaluation

The evaluation used in this experiment is non-traditional and it is based on the subtopic relevance retrieval evaluation procedure proposed in (Zhai et al., 2003). In the evaluation, they used The Financial Times of London (1991-1994) collection. The collection contains 210,158 documents totalling about 500MB with the average document length of 400 words (Zhai et al., 2003). The topics and the relevance judgment data are taken from the interactive track of the TREC-6, TREC-7 and TREC-8, which consists of 20 topics in total. In order to facilitate the interactive retrieval task, a set of instances (aspects of the topic) is defined for each topic. The instances indicate the aspects in which the retrieved documents should be judged. As a result, the retrieved documents are judged according to these aspects of the topic. For instance the topic 392i with the title '*robotics*' has a set of instances such as '*spot-welding robotics*', '*controlling inventory – storage devices*' and so on (Zhai et al., 2003). The retrieved document is judged based on its relevance to these instances. A document can be relevant to some of the instances of the topic. In this subtopic relevance retrieval evaluation, these instances are considered as the subtopics for the topic as proposed in (Zhai et al., 2003). Each topic has a set of subtopics ranging from 7 to 56 subtopics.

The only difference of the evaluation in this chapter and the one conducted in (Zhai et al., 2003) is that the effectiveness of the ranking produced by an IR system is measured for the top-*n* documents instead of the entire ranking produced by the system. In this evaluation, we are not interested in measuring the overall retrieval performance due to the fact that most post-retrieval browsing occurs only for the top-*n* documents. One of the reasons is that the number of documents that a user may assess during the browsing is small and it is unlikely that the user assesses the entire ranking of the documents.

6.3.1 Evaluation measure

Since it is a non-standard evaluation of the retrieval performance, a different evaluation measure is required for the subtopic relevance retrieval experiment. In (Zhai et al., 2003), the authors propose the *s_recall* and the *s_precision* measures for the subtopic relevance retrieval evaluation. These measures are calculated based on the assumption that a 100% recall of the subtopics for a given topic can be found within the entire ranking of the documents. Since we are only interested in the top-*n* documents, and not the entire ranking, we may not achieve 100% subtopic recall in most cases. Therefore, a slight modification on the evaluation measures is made for the purpose of this experiment. The subtopic recall at rank q for n documents is measured by:

$$s\_recall_{qn} = \frac{\left| \bigcup_{i=1}^{q} subtopics(d_i) \right|}{\left| \bigcup_{i=1}^{n} subtopics(d_i) \right|} \tag{6-8}$$

where $subtopics(d_i)$ are the set of relevant subtopics for document d_i. If $q = n$, then $s\_recall_{qn} = 1.0$. If S is some system that produces a ranking of documents and y is a subtopic recall level, $0 \le y \le 1$, $\text{minRank}(S, y)$ is defined to be the minimal rank q at which the ranking produced by S has s_recall_{qn} y. Thus, subtopic precision at subtopic recall y is defined as:

$$s\_precision_y = \frac{\text{minRank}(S_{opt}, y)}{\text{minRank}(S, y)} \tag{6-9}$$

where S_{opt} is a system that produces the *optimal* ranking that obtains *y*, such that the rank is the smallest possible *q*. A detailed discussion on these evaluation measures can be found in (Zhai et al., 2003). In our evaluation, the *optimal* ranking is defined by ranking the top-*n* documents to achieve the highest s_recall_{qn} at each q i.e. a greedy approach that iteratively selects the remaining document with the most number of relevant subtopics at each rank.

The subtopic precision score is computed for each defined subtopic recall level, from zero to one, inclusive, such as at $y = (0.0,0.2,0.4,0.6,0.8,1.0)$. For instance, a subtopic precision of 0.5 could be measured on a subtopic recall level one (a 100% subtopic recall). The score indicates that for a given ranking, the number of documents need to be assessed to achieve a 100% subtopic recall is double of the ideal ranking. This means that the effort that needs to be spent to browse a given ranking of documents is double the ideal ranking. When the subtopic precision is one, the quality of a given ranking is as good as the ideal ranking. A higher subtopic precision is more desirable for this evaluation.

6.4 Experimental Results

First, the effectiveness of the baseline ranking method that is to rank the retrieved documents based on their RSVs, is evaluated in Section 6.4.1. Then, the effectiveness of re-ranking the top-*n* documents based on their novelty scores given by the novelty models is investigated in Section 6.4.2. After that, the effectiveness of using the MMR scores, a linear combination of the relevance scores (RSV values) and the novelty scores, to re-rank the top-*n* documents will be discussed in Section 6.4.3. Finally, the effectiveness of using the intertemporal choice browsing model to re-rank the top-*n* documents based on the novelty scores and the MMR scores is presented in Section 6.4.4.

6.4.1 Baseline

The aim of this evaluation is to prove the hypothesis that the conventional ranking method is less effective than the MMR ranking method for the subtopic relevance retrieval application. It is based on the assumption that the effectiveness of the subtopic retrieval application depends on the sensitivity of the ranking toward novelty, and the MMR ranking method incorporates this notion in its ranking. For this evaluation, the Terrier retrieval system (Ounis et al., 2005) is used as the baseline to produce the initial ranking of documents for the queries. The relevance values of all documents in this experiment are also based on the relevance values produced by Terrier.

In this evaluation, default configuration of Terrier is used where the parameter *c* is not optimised and is set to its default value, 1.0. Moreover, the number of the top retrieved documents, the top-*n*, to be re-ranked is set to 10. Furthermore, in the experiments conducted previously, the *InL2* weighting model of the Terrier system appeared to be

the best model for this evaluation and this test collection. Therefore, the baseline Terrier system is using the *InL2* model with default configuration.

Figure 6-2 shows the subtopic precision scores of the baseline at different subtopic recall levels. At subtopic recall 0, the subtopic precision score is usually 1. The subtopic precision score gradually decreases as the subtopic recall level increases. At subtopic recall 0.5, when half of the subtopics for a given topic have been found, the subtopic precision score of the baseline is about 0.5518. It means that the number of documents need to be assessed following the ranking produced by the baseline system to arrive at subtopic level 0.5 is twice the number of documents need to be assessed for the optimal ranking. Moreover, the subtopic precision score does not reduce much to arrive at subtopic recall 1, where the score is 0.5110.

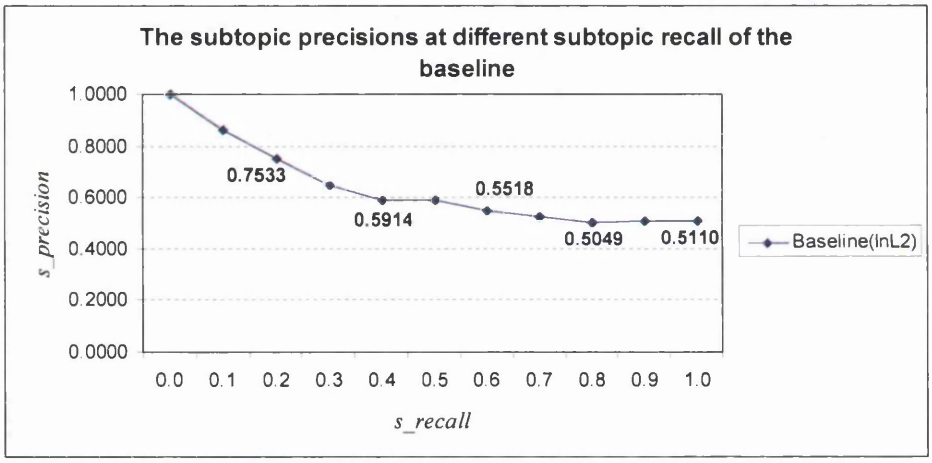


Figure 6-2: The subtopic precision scores of the baseline at each subtopic recall level

Note that the baseline uses only the relevance score of the documents to produce the ranking. Next, the performance of the system that uses only the novelty score to rank the documents is evaluated in Section 6.4.2. After that, the Maximal Marginal Relevance (MMR) technique that combines the relevance score and the novelty score of documents in the ranking is evaluated in Section 6.4.3. Finally, the experiment will evaluate the effectiveness of the intertemporal choice browsing model in producing an optimal ranking by using the novelty score and also the MMR score in Section 6.4.4.

6.4.2 Novelty model

In the case of the MMR ranking method, the novelty scores are incorporated into the ranking together with the relevance scores. In order to apply the MMR ranking method, a good novelty model needs to be chosen for this application. Therefore, in the

ollowing experiments, the effectiveness re-ranking the top-*n* documents by using only he novelty scores is investigated. The novelty scores are computed by eight novelty nodels, four distance-based models and four set-based models.

6.4.2.1 Distance-based novelty models

n essence, the distance-based novelty models measure the novelty of a document by omputing the distance (or the similarity) between the document and all other previous locuments in a browsing session. The novelty score is estimated based on either its verage distance (or similarity) or its minimum distance (or maximum similarity). The our distance-based novelty models investigated are the *average cosine distance* (DN1), he *minimum cosine distance* (DN2), the *average Kullback-Liebler similarity* (DN3) and he *maximum Kullback-Leibler similarity* (DN4). The description of the models can be ound in Section 6.2.2.1.

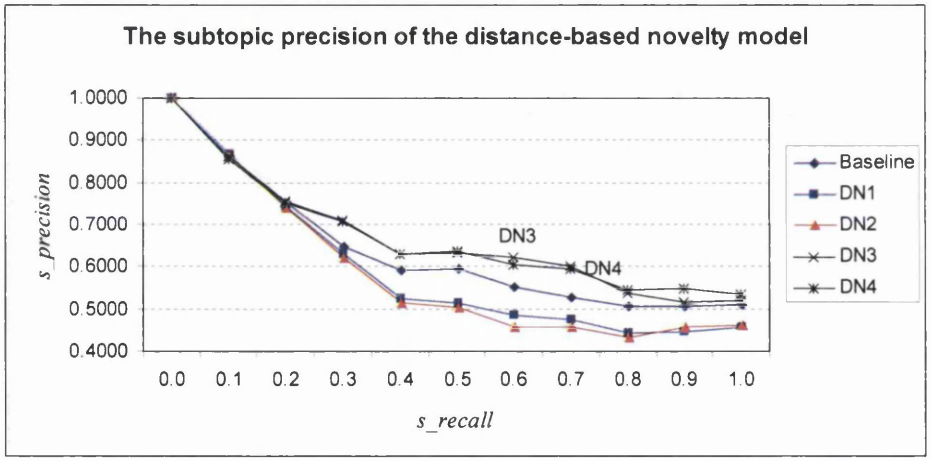


Figure 6-3: The subtopic precision for the distance-based novelty models

Based on Figure 6-3, the subtopic precisions of all models including the baseline are similar when the subtopic recall is between 0 and 0.2. The scenario changes when the subtopic recall is 0.3 and above, whereby DN3 and DN4 is consistently better that the baseline and DN1 and DN2 are consistently worse than the baseline. The subtopic precision of DN3 and DN4 is higher than the subtopic precision of the baseline in almost all level of subtopic recall.

Table 6-3 shows the subtopic precision of the novelty models and its percentage difference to the subtopic precision of the baseline. Based on the average percentage difference, it is obvious that only DN3 and DN4 are better than the baseline, with the average percentage difference of 5.48% and 5.89%, respectively. Moreover, based on

the results, it is learnt that the *Kullback-Leibler* based novelty model is better than the *Cosine distance* based novelty model. However, the assumption that the minimum distance (or maximum similarity) novelty models are better than the average distance (or similarity) novelty model is incorrect since DN2 is worse than DN1 and DN4 is not significantly better than DN3 (p=0.779, paired t-test). Based on this result, DN4 will be chosen as one of the novelty models for the remaining experiments in this chapter.

Table 6-3: The percentage difference of the subtopic precision for the distance-based novelty model

s_recall	Baseline	DN1	Δ %	DN2	Δ %	DN3	Δ %	DN4	Δ %
0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0
0.1	0.862	0.866	0.53	0.864	0.26	0.862	0.0	0.856	-0.63
0.2	0.753	0.741	-1.60	0.741	-1.70	0.758	0.55	0.752	-0.17
0.3	0.647	0.630	-2.58	0.620	-4.15	0.712	10.05	0.706	9.21
0.4	0.591	0.522	-1.68	0.512	-13.51	0.628	6.24	0.630	6.59
0.5	0.592	0.512	-13.65	0.504	-14.91	0.634	7.05	0.636	7.41
0.6	0.552	0.484	-12.26	0.458	-16.99	0.622	12.65	0.604	9.41
0.7	0.528	0.474	-10.23	0.457	-13.39	0.603	14.09	0.594	12.55
0.8	0.505	0.441	-12.68	0.431	-14.68	0.537	6.35	0.544	7.77
0.9	0.506	0.447	-11.78	0.457	-9.70	0.515	1.67	0.548	8.23
1.0	0.511	0.458	-10.39	0.460	-10.01	0.519	1.61	0.533	4.39
Average			**-7.85**		**-8.98**		**5.48**		**5.89**

6.4.2.2 Set-based novelty models

There are four set-based novelty models being investigated in this evaluation. The models are based on the *new word* measure (SN1), the *normalized new word* measure (SN2), the *overlap* measure (SN3) and the *set difference* measure (SN4). The description of the models can be found in Section 6.2.2.2.

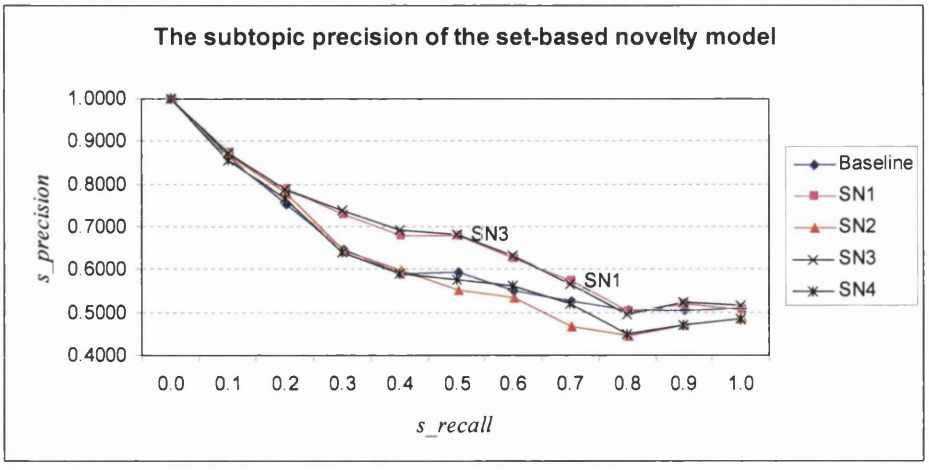


Figure 6-4: The subtopic precision for the set-based novelty models

Table 6-4: The percentage difference of the subtopic precision for the set-based novelty models

s_recall	Baseline	SN1	Δ %	SN2	Δ %	SN3	Δ %	SN4	Δ %
0.0	1.0	1.0	0.0	1.0	0.00	1.000	0.00	1.000	0.00
0.1	0.862	0.872	1.26	0.864	0.29	0.872	1.26	0.856	-0.68
0.2	0.753	0.789	4.76	0.777	3.10	0.789	4.76	0.765	1.61
0.3	0.647	0.727	12.37	0.643	-0.52	0.739	14.30	0.640	-1.09
0.4	0.591	0.680	14.98	0.597	1.03	0.693	17.09	0.590	-0.30
0.5	0.592	0.678	14.37	0.553	-6.71	0.683	15.21	0.575	-2.85
0.6	0.552	0.626	13.41	0.536	-2.92	0.633	14.77	0.563	2.04
0.7	0.528	0.573	8.41	0.468	-11.45	0.565	7.07	0.520	-1.61
0.8	0.505	0.501	-0.73	0.445	-11.93	0.496	-1.78	0.448	-11.21
0.9	0.506	0.520	2.73	0.470	-7.10	0.523	3.19	0.471	-6.90
1.0	0.511	0.504	-1.31	0.484	-5.33	0.517	1.16	0.486	-4.90
Average			**6.38**		**-3.78**		**7.00**		**-2.35**

Based on Figure 6-4, the subtopic precision for SN1 and SN3 is clearly better than the subtopic precision for the baseline when the subtopic recall level is between 0.2 and 0.8. The performance of SN2 and SN4, however, is not better than the baseline at almost all subtopic recall levels. The pattern is more obvious in Table 6-4, where the average percentage difference of the subtopic precision between the novelty models and the baseline shows that SN1 and SN3 are better that the baseline, while SN2 and SN4 are not better than the baseline. SN1 and SN3 are both based on the same principle, in which the novelty score is measured based on the number of new words in a document with respect to the other documents. It is believed that different words can be used to discriminate between different subtopics for a given topic. Therefore, the models based on the new word counts will work well as a novelty measure.

Based on the results so far, it is learnt that only four novelty models produce a better ranking of documents based on this subtopic relevance retrieval evaluation, which are DN3, DN4, SN1 and SN3.

6.4.3 Maximal Marginal Relevance (MMR) ranking

Based on the observation of the effectiveness of different novelty models in Section 6.4.2, three models, DN4, SN1 and SN3, are chosen as the novelty models for the remaining experiments in this chapter. The models show an improvement on the subtopic precision of the baseline. It is assumed that the performance of these models could be further improved by using the MMR ranking method.

In essence, the MMR ranking method uses the combination of the relevance score and the novelty score as the final score for a document in the ranking. In this evaluation, a linear combination model is used to combine the scores. The range of the values for

relevance scores and novelty scores are not necessarily the same. Therefore, for the purpose of the combination, the respective scores are normalized to the maximum value of the scores in the set. The MMR score is computed by the following equation.

$$mmr(d_i) = \mu \times relevance(d_i) + (1-\mu) \times novelty(d_i) \quad (6\text{-}10)$$

where $relevance(d_i)$ is the relevance score for document d_i and $novelty(d_i)$ is the novelty score for document d_i while μ is the combination parameter. The value of μ is between 0.0 and 1.0, inclusive, where only the relevance score is used when $\mu = 1.0$ and only the novelty score is used when $\mu = 0.0$.

6.4.3.1 The effectiveness of the MMR ranking method

In order to evaluate the effectiveness of the MMR ranking method, DN4, SN1 or SN3 is used as the novelty models to compute the novelty score of a document. The novelty score is linearly combined with the relevance score of the document assigned by the baseline retrieval system to produce an MMR score for the document. The top-n retrieved documents are re-ranked based on their MMR score to produce a ranking that is sensitive for both relevance and also novelty. Such a ranking is more suitable for the subtopic relevance retrieval application.

Based on Equation (6-10), the combination of the relevance and novelty score depends on the value of the combination parameter, μ, where $0 \leq \mu \leq 1$. Figure 6-5 shows the graphs for the subtopic precision score of the MMR ranking method when DN4, SN1 and SN3 are used as the novelty scores. The graphs are plotted for different μ. In addition, the subtopic precision score reported in the graph is the score at subtopic recall level 1.0 (when all subtopics are retrieved).

Based on Figure 6-5, it is obvious that DN4 is the best novelty model to rank the documents based only on the novelty scores (when $\mu = 0.0$). However, the combination of the novelty score produced by the model with the relevance score of a document (as the MMR score for the document) does not improve the subtopic precision of the ranking produced. In fact, the subtopic precision of the ranking may be worse than the subtopic precision of the baseline (when $\mu = 1.0$).

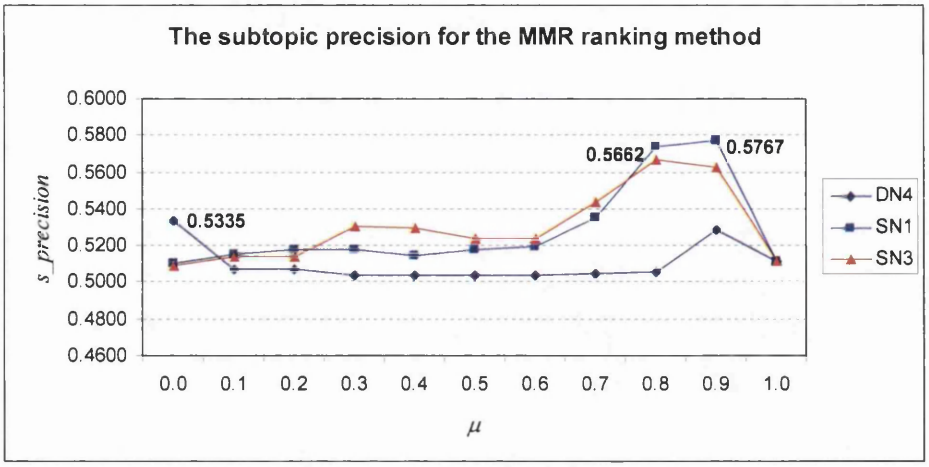


Figure 6-5: The subtopic precision of the MMR ranking method

On the other hand, the benefit of the MMR ranking method is very obvious when SN1 or SN3 is used as the novelty model. Based on Figure 6-5, an improvement of the subtopic precision is observed MMR score is used with either SN1 or SN3. It is also learnt that as the value of the combination parameter increases up to 0.8, in which the strength of the relevance value increases, the performance of the system increases. SN1 reaches its peak performance when $\mu = 0.9$, while SN3 reaches its peak performance $\mu = 0.8$. The best subtopic precision of 0.5767 is achieved when SN1 is used and $\mu = 0.9$.

Based on the results, it can also be observed that the combination of the relevance and novelty score may reduce the performance of the system, for DN4 in particular. In this experiment, the relevance values and the novelty values are normalised before they are combined. In this case, they are normalised to the maximum values to ensure that all relevance and novelty values are between 0 and 1. Therefore, it is believed that such a normalisation technique might only work for some novelty model. Therefore, further experiments should be conducted by using different normalisation techniques to actually verify this statement.

6.4.4 The intertemporal choice browsing model

So far, we have discussed the problem of browsing through the top-n retrieved documents in the context of the subtopic relevance retrieval problem. In this context, the importance of the next documents to be assessed depends on the novel information it contains. In the previous experiments, we have seen that the effectiveness of the system can be improved by using novelty scores, as opposed to the relevance scores (or

the RSVs), to re-rank the top-*n* retrieved documents for some novelty models investigated in this evaluation. Furthermore, by combining the novelty scores and the relevance scores in the MMR ranking method, a further improvement of the subtopic precision score for the system can be achieved.

The main assumption behind the MMR ranking method is that it provides a ranking of documents that encourages browsing. However, due to the fact that the method adopts a greedy approach for the document selection in the ranking, there is a possibility that the ranking it produces is not optimal. A more exhaustive approach, which considers all possible rankings that can be generated from the set of documents before choosing the best ranking, could be a better method. In this section, the evaluation of such method is discussed. Each possible ranking that can be generated from a set of documents is assigned a score that indicates its importance to the user. The best ranking of documents is the one with the highest score.

The intertemporal choice browsing model is used to assign a score to each ranking. The model assigns a score to an ordered of documents based on the importance of the sequence to the user. It applies an appropriate discount function as a trade-off for user satisfaction and the length of browsing. In this experiment, three discount functions, HARVEY, SAMUELSON and MAZUR are evaluated. SAMUELSON represents an exponential discount function, while HARVEY and MAZUR represent hyperbolic discount functions. The details of the intertemporal choice browsing model can be found in Section 6.2.3 and in Chapter 3.

Table 6-5: The ranges of the parameter's value for the discount functions

Discount function	*Parameter*	*Range of values*
SAMUELSON	r	0.1 to 8.0
MAZUR	s	0.25 to 5.0
HARVEY	h	0.1 to 8.0

The discount functions of the intertemporal choice browsing model are parametric. Therefore, the parameter of the function needs to be tuned for the best performance. Table 6-5 shows the range of values of the parameter for each discount function to be used in this experiment.

5.4.4.1 Intertemporal choice browsing with novelty

The intertemporal choice browsing model consists of a discount function and a value function as described in Section 6.2.3. The value function transforms the objective value of a document into a subjective value. In this experiment, the objective value of a document is computed based on the novelty model. Based on the results from the experiments in Section 6.4.2, we have observed that the DN4, the SN1 and the SN3 are reasonable novelty models for this application. Therefore, the objective value of a document is computed by the DN4, the SN1 and the SN3 novelty models.

Figure 6-6, Figure 6-7 and Figure 6-8 depict the subtopic precision score for the intertemporal choice browsing model with HARVEY, SAMUELSON and MAZUR as the discount functions, respectively. The subtopic precision score is plotted against different values of *h*, *r* and *s*.

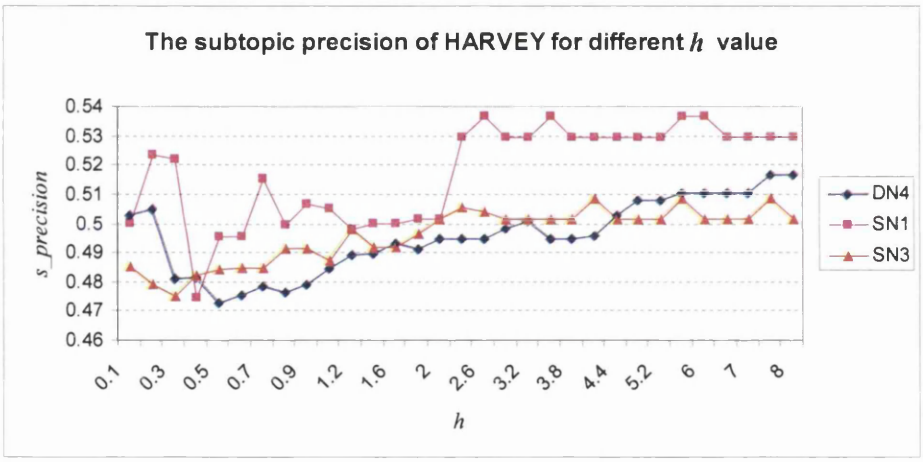


Figure 6-6: The subtopic precision score of novelty models with HARVEY

In this experiment, different parameters values are used to find the optimal setting for the discount function. It is also aimed to discover patterns on the relationship between the parameters values and the effectiveness of the intertemporal choice browsing model. According to the graphs in Figure 6-6, Figure 6-7 and Figure 6-8, there is no obvious general pattern relating the parameters values and the effectiveness of the intertemporal choice browsing model. Furthermore, there is no absolute optimal setting for the model.

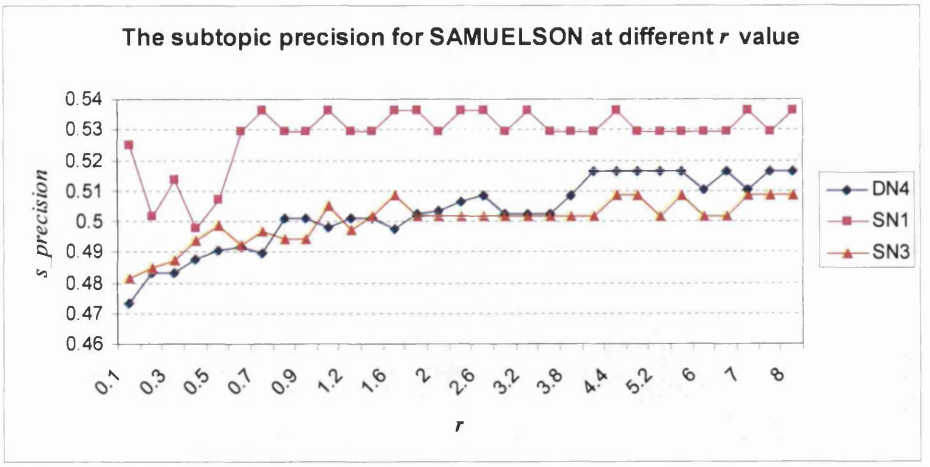


Figure 6-7: The subtopic precision score of novelty models with SAMUELSON

However, for the intertemporal choice browsing model with DN4 and SN3, a gradual increasing pattern of the subtopic precision score can be seen as the value of the parameter increases. Such pattern is more obvious for the intertemporal choice browsing model with HARVEY and SAMUELSON. For the SN1 model, the effectiveness of the system is rather low when the value of the parameter is smaller before it rises to a slightly consistent value as the parameter value increases. Such pattern can be seen for the intertemporal choice browsing model with HARVEY and SAMUELSON.

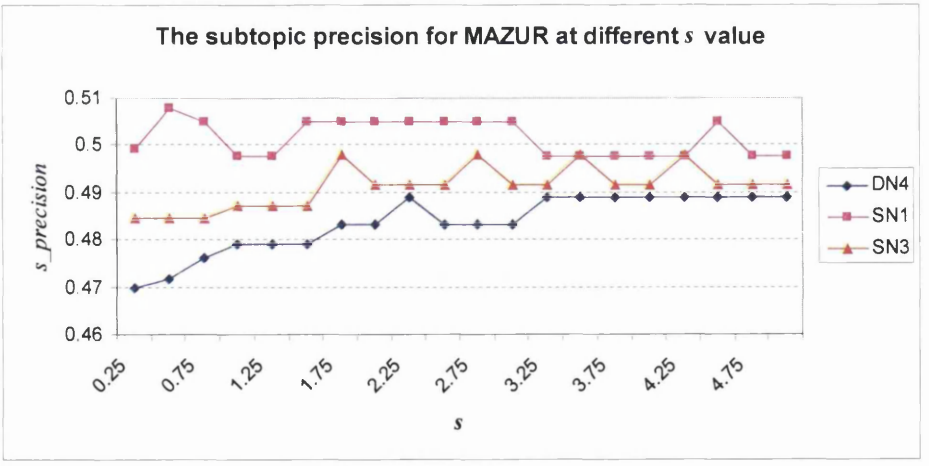


Figure 6-8: The subtopic precision score with MAZUR

As depicted in Figure 6-8, the subtopic precision scores for MAZUR is rather consistent with DN4 and SN3 showing a slight improvement as the parameter value increases. Overall, it is hard to determine the best configuration for the discount functions. There is no common pattern to predict the behaviour. Therefore, the configuration that produces the best subtopic precision score is chosen for this experiment. Table 6-6

shows the best subtopic precision score for each discount function as compared to the two baselines, using relevance and novelty only.

Table 6-6: The best subtopic precision score for each discount function with its baselines

Novelty Model	*Baseline*	*Novelty Only*	*Intertemporal Choice Browsing Model (Best score)*		
			HARVEY	**SAMUELSON**	**MAZUR**
DN4	0.5110	0.5335	0.5165	0.5165	0.4888
SN1	0.5110	0.5106	*0.5362*	*0.5362*	0.5077
SN3	0.5110	0.5086	0.5085	0.5085	0.4979

First, let us look at the results for the DN4 novelty model as depicted in Table 6-6. For the intertemporal choice browsing model with HARVEY and SAMUELSON as the discount function, an increase of about 1% of the subtopic precision can be seen as compared to the baseline. However, the best score for MAZUR is about 4.3% less than the baseline. However, the improved score for HARVEY and SAMUELSON is still about 3.2% less than the ranking produced based on DN4 only.

As for the SN1 novelty model, the subtopic precision score for HARVEY and SAMUELSON is 4.9% better than the baseline and is 5% better than the ranking based on the SN1 only. Similar to the DN4 novelty model, the score for MAZUR is not improved, at 0.7% less than the baseline. However, in the case of the SN3 novelty model, there is no improvement of the subtopic precision score against the baseline and against the ranking based on SN3 only.

Interestingly, there is a very small improvement of the subtopic precision score for the best configured intertemporal choice browsing model and the best novelty only ranking model. The score for the SN1 model with HARVEY or SAMUELSON is 0.5% better than the ranking with DN4 only. Furthermore, the best subtopic precision scores for HARVEY and SAMUELSON are similar for different novelty models.

6.4.4.2 The intertemporal choice browsing with MMR

In Section 6.4.3, it has been shown that the subtopic precision score can be improved by combining the relevance score and the novelty score in the MMR ranking method for the SN1 and the SN4 novelty models. In this section, the effect of using the MMR score for the intertemporal choice browsing model on the effectiveness of the system is investigated. According to the results presented in Section 6.4.3, a good value of μ is

0.9 for DN4, 0.9 for SN1 and 0.8 for SN3 novelty model. Therefore, these values will be used in this experiment.

The experiment is conducted with different values for the parameter of the discount functions similar to the experiment in Section 6.4.4.1. Similarly, there is no obvious pattern relating the different values for the parameters and the performance of the system. Due to the fact, it is difficult to find the optimal setting for the parameters. Therefore, the discussion will be focused to the best subtopic precision achievable from the experiment as summarised in Table 6-7.

Table 6-7: The best subtopic precision score with MMR and its baselines

Novelty Model	*Baseline*	*MMR Baseline*	*Intertemporal Choice Browsing Model (MMR)*		
			HARVEY	SAMUELSON	MAZUR
DN4	0.5110	0.5282	0.5069	0.5102	0.4522
SN1	0.5110	0.5767	0.5445	0.5326	0.5128
SN3	0.5110	0.5662	0.5113	0.4864	0.4356

The performance of the intertemporal choice browsing model with the HARVEY, SAMUELSON and MAZUR discount functions will be discussed for each novelty model used for the MMR score. For DN4, the intertemporal choice browsing model does not improve the performance of the baseline (the InL2) or the MMR baseline. As for SN1, an increase of 6.6%, 4.2% and 0.4% as compared to the baseline is observed for HARVEY, SAMUELSON and MAZUR, respectively. However, the improvement shown is still less than the MMR baseline. As for the SN3 model, the performance of the system is generally not better than the baseline for all discount functions, while the subtopic precision for HARVEY is as good as the baseline. Similarly, its performance is worse than the MMR baseline. Furthermore, the best configuration of the intertemporal choice browsing model fails to improve the best score for the MMR baseline.

6.4.4.3 The intertemporal choice browsing without its value function

It is unfortunate to learn that the intertemporal choice browsing model is unable to improve the subtopic precision score for the novelty only baseline or the MMR baseline as presented in Section 6.4.4.1 and Section 6.4.4.2, respectively. Does the value function component of the model affect its performance?

The value function is applied to the objective value of the document to model the *risk aversion* of an individual in the decision making scenario. The monotonic increasing concave function for positive values (or the monotonic decreasing convex function for negative values) transforms the objective value of the document into a subjective value, which represents the risk aversion scenario (Kahneman and Tversky, 1979). However, such behaviour may not necessarily occur in the context of browsing. Therefore, in order to observe the effect of the value function on the performance of the intertemporal choice browsing model, the function will be removed from the model. Therefore, in the model, the discounting is applied to the objective value of the documents (the values computed by the novelty models or the MMR model) rather than the subjective value of the documents. Equation (6-7) is revised as follows:

$$U_k = U(u_1, t_1; \ldots; u_n, t_n) = \sum_{i=1}^{n} u_i \times \phi(t_i) \tag{6-11}$$

where u_i is the value of the document and $\phi(t_i)$ is the discount function applied to t_i. As we can observe from Equation (6-11), the discounting is applied directly to the value of the documents.

The aim of this experiment is to compare the performance of the intertemporal choice model with and without the value function. Again, the model is configured to its best performance as observed from the experiments in Section 6.4.4.1 and Section 6.4.4.2. Table 6-8 and Table 6-9 show the subtopic precision score for the intertemporal choice browsing model with the value function (with VF) and without the value function (no VF). The novelty score is used in Table 6-8 while the MMR score is used in Table 6-9.

Table 6-8: The subtopic precision with and without the value function (novelty score)

Novelty Model	*HARVEY*		*SAMUELSON*		*MAZUR*	
	With VF	**No VF**	**With VF**	**No VF**	**With VF**	**No VF**
DN4	0.5165	0.5103	0.5165	0.5103	0.4888	0.4847
SN1	0.5362	0.5291	0.5362	0.5362	0.5077	0.5291
SN3	0.5085	0.5014	0.5085	0.5014	0.4979	0.4830

Table 6-9: The subtopic precision with and without the value function (MMR score)

Novelty Model	*HARVEY*		*SAMUELSON*		*MAZUR*	
	With VF	**No VF**	**With VF**	**No VF**	**With VF**	**No VF**
DN4	0.5069	0.5119	0.5102	0.5102	0.4522	0.4421
SN1	0.5445	0.5362	0.5326	0.5362	0.5128	0.5142
SN3	0.5113	0.5042	0.4864	0.4550	0.4356	0.4312

Based on these results, the removal of the value function from the intertemporal choice browsing model does not improve the effectiveness of the system. This can be observed in all configurations except for MAZUR with the SN1 novelty model in Table 6-8 and for HARVEY with the DN4 novelty model as well as for MAZUR with the SN1 novelty model in Table 6-9.

6.4.4.4 The performance of the intertemporal choice browsing model at different subtopic recall levels

So far, the effectiveness of an IR system for post retrieval browsing is measured based on the subtopic precision when the subtopic recall is 1.0. Such measurement assumes that the user will only be satisfied when all the relevant subtopics in the top-*n* documents have been found. It is a reasonable measure to compare the effectiveness of different systems. However, it is possible that the user will be satisfied once some of the relevant subtopics are discovered. In such cases, the effectiveness of an IR system can be measured by the subtopic precision at different subtopic recall levels.

Figure 6-9 and Figure 6-10 shows the subtopic precision scores at different subtopic recall levels for the system that uses the SN1 novelty score and the MMR score, respectively. The SN1 novelty model is chosen for this discussion since the model appears to be the best in most experiments in this chapter.

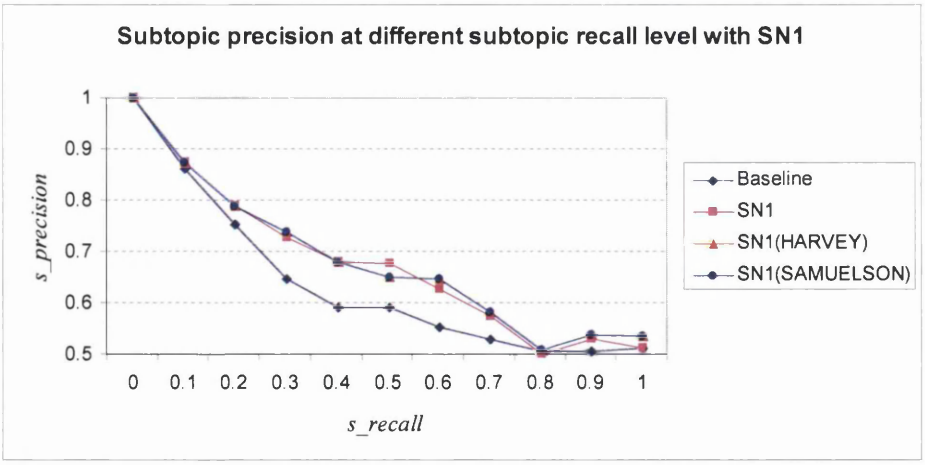


Figure 6-9: The subtopic precision score with SN1 at different subtopic recall level

Based on Figure 6-9, it is obvious that the benefit of using the novelty scores in the ranking to the user is realized when he/she is looking for more than 20% of the relevant subtopics in the top-*n* documents as compared to the baseline InL2 ranking. The pattern is consistent for higher subtopic recalls except when it is 0.8. The performance for SN1

and the intertemporal choice browsing model with SN1 are rather similar for all subtopic recall levels.

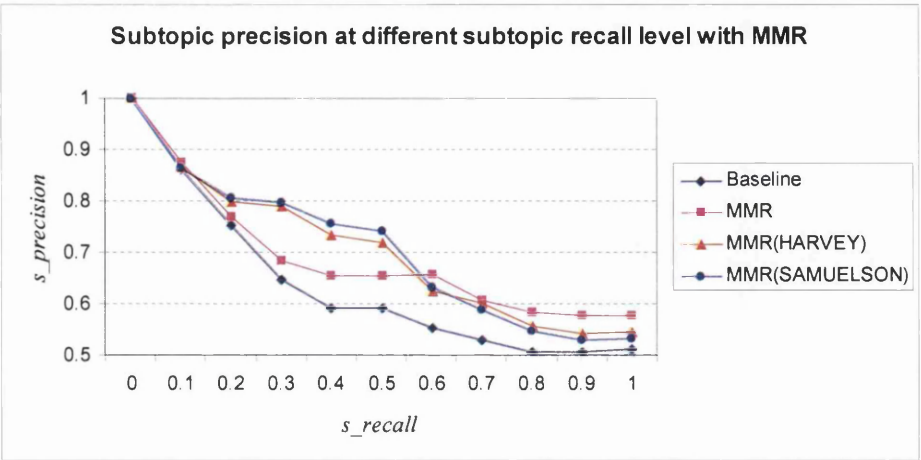


Figure 6-10: The subtopic precision score with MMR at different subtopic recall level

As shown in Figure 6-10, the performance of the system that uses the MMR score is much better than the baseline for all subtopic recall levels. In this case, the intertemporal choice browsing model outperforms the MMR baseline model when the subtopic recall level is between 0.2 and 0.5, inclusively. At the same range, SAMUELSON is a slightly better discount function than HARVEY. For higher subtopic recall levels, the baseline MMR with SN1 as the novelty model is a better system.

6.5 Discussion

In this chapter, the effectiveness of the ranking of documents is investigated for post retrieval browsing. With the assumption that the ranking could be a reasonable guidance for the user to decide on his/her browsing strategy, it is important that the ranking is optimal. A few techniques to produce a better ranking for documents were described (see Section 6.2) and the results of the evaluations were presented (see Section 6.4). In this section, the evaluation results will be discussed in the context of the sub-optimality of the Probability Ranking Principle (PRP) based ranking, the novelty as the dependent relevance and the effect of the intertemporal choice browsing model.

6.5.1 Sub-optimality of the PRP based ranking

First, the effectiveness of the conventional ranking system that is based on the PRP for the post retrieval browsing has been investigated. The ranking is compared to a number of alternative methods such as the Maximal Marginal Relevance (MMR) ranking

method and the intertemporal choice browsing model. The main assumption is that a sequential assessment of documents in browsing creates the situation where the relevance of a document may be judged based on its relevancy to the topic as well as the contents of the documents already seen so far. Therefore, to rank the retrieved documents based solely on their relevance values (based on the PRP) may not be the optimal way for post retrieval browsing. By evaluating the hypothesis based on the subtopic relevance retrieval application, it has been shown that the conventional ranking system is less effective for post retrieval browsing.

Based on the experimental results, it has been discovered that the effectiveness of a ranking for browsing can be improved when a dependent relevance value is used to rank the documents. For this application, the dependent relevance value is given by the novelty score computed by a number of different novelty models. The results from the experiments have shown that the use of the novelty score individually (see Section 6.4.2) or in combination with the relevance score through MMR (see Section 6.4.3) in the ranking could improve the effectiveness of the ranking in many cases. The use of the scores in the best configured intertemporal choice browsing model helps to improve the effectiveness of the system against the conventional ranking system (see Section 6.4.4). Therefore, it is safe to say that the conventional ranking system is not optimal for post-retrieval browsing.

Such discovery proves the claims made by a number of researchers on the sub-optimality of the PRP based ranking when a dependent relevance is assumed (Robertson, 1977, Gordon and Lenk, 1991, Gordon and Lenk, 1992). In (Gordon and Lenk, 1992), the authors show that the ranking becomes sub optimal when the documents in the ranking are not assessed independently, in which the relevance of a document now depends on other documents already assessed by the user. The same argument applies for browsing.

The use of novelty to model dependent relevance is very appropriate in the context of the subtopic relevance retrieval application. Dependent relevance can also be modelled differently. In the *indirect retrieval method*, Goffman suggested that relevance of a document could depend on the document he/she just read (Croft and van Rijsbergen, 1976).

6.5.2 Novelty as dependent relevance

In this chapter, the concept of dependent relevance is modelled as novelty. The value of a document is estimated based on its novelty against the other documents. The ranking that is based on dependent relevance is more effective for post-retrieval browsing compared to the conventional ranking based on independent relevance, as evaluated in the context of the subtopic relevance retrieval application.

Although the concept of dependent relevance has been successfully modelled as novelty, the performance of some novelty models is rather disappointing. Out of the eight novelty models, only four models successfully improve the effectiveness of the system. The models have been applied in many novelty detection applications elsewhere (Zhang et al., 2002, Allan et al., 2003, Zhai et al., 2003, Gabrilovich et al., 2004).

The combination of the novelty scores and the relevance scores in the ranking improves the effectiveness of the system. In some cases, the effectiveness of the system is better than using the novelty score or the relevance score individually (see Section 6.4.3). Based on the experimental results, there is an interesting pattern in term of the strength of the combination. It seems that the best performance is achieved when the relevance score is given more importance although the use of the novelty score is better than the relevance score, individually.

6.5.3 The impact of the intertemporal choice browsing

The fundamental idea of the intertemporal choice model is the trade-off between getting better outcomes and the delay for getting the outcomes (Read, 2004). Based on the idea, the intertemporal choice browsing model is proposed to model the browsing behaviour of the user. In this case, the trade-off is between getting a better and more relevant document and the cost of browsing (or the length of browsing). With the assumption that the model could assign a value reflecting the importance of a browsing strategy to the user, an optimal browsing strategy could be found.

The intertemporal choice browsing model discussed in this chapter has two important characteristics. First, the model assumes a positive time preference (Loewenstein and Prelec, 1993). It is based on the assumption that one of the objectives for the subtopic relevance retrieval application is to reduce the cost of browsing (the number of

documents needs to be browsed). Thus, a user is assumed to be impatient in his/her information seeking task. Second, the model assumes a dependent relevance value for the objective value of the documents, which is modelled by novelty. This is because the novelty model is more appropriate for browsing, especially in the context of the subtopic relevance retrieval application, as shown in the results of the experiments.

Is the model able to predict a better browsing strategy? It could improve the effectiveness of the system to a certain extent. For example, in many cases, the best configuration of the model could improve the effectiveness of the conventional ranking system. However, the use of the novelty score in the model could be the reason for this to happen, since it has been shown that the novelty-based model could improves the effectiveness of the system as well. In fact, only in very selective cases does the model predict a better browsing strategy as compared to a simple ranking based on the novelty score. In fact, the model is still unable to predict a better browsing strategy compared to the MMR approach.

Furthermore, tuning the parameter for the discount function is troublesome. Therefore, it is difficult to set the model to its best performance. Based on the experimental results, there is no obvious pattern on the relationship between the performance of the model and the increase of parameter value for the discount function. However, by setting the parameter value to a higher value, it is likely that the model will perform better (see Section 6.4.4.1). It should be the case since the value of a document will be discounted heavily as the length of browsing increases.

One interesting observation is that the performance of SAMUELSON and HARVEY as the discount functions is comparable, while they mostly outperform MAZUR. The fact that SAMUELSON is an exponential function and HARVEY is a hyperbolic function does not make any difference in performance in this context. This finding contradicts the hypothesis that suggests a hyperbolic function is better than an exponential function. It is based on the results from many studies in the area of economy (Loewenstein and Prelec, 1992, Cairns and van der Pol, 2000, Frederick et al., 2002, Lazaro et al., 2002, Read, 2004).

The purpose of the value function component in the intertemporal choice browsing model is to model the *risk aversion* (Kahneman and Tversky, 1979) behaviour of an

individual in the decision making process (Loewenstein and Prelec, 1992). Usually, risk aversion is discussed in the context of monetary outcomes, where it is related to an individual's fear of losing. However, in this chapter, the outcomes are documents and it is possible that such behaviour may not be evident in browsing. Therefore in Section 6.4.4.3, the possibility of removing the value function from the model was discussed with the intention to improve the quality of the prediction by the system. However, it seems that removing the value function will not improve the quality of the prediction. Furthermore, the experimental results suggest that the effectiveness of the system may reduce as a result of the elimination. In this case, more evidence is needed by analysing the behaviour of the real users.

Finally, the novelty based systems (the novelty based ranking, the MMR based ranking and the intertemporal choice browsing model) are more effective than the conventional ranking system even when the user is satisfied by finding information on only some and not all subtopics in the top-*n* documents (see Section 6.4.4.4). Therefore, the system will still be helpful, if the user is only looking for some subtopics.

6.6 Chapter Summary

This chapter attempted to find a better way to re-rank the top-*n* documents for post retrieval browsing. The main hypothesis is that the conventional ranking based on the PRP may not be optimal, since it does not take into consideration the dependent relevance assumption in the ranking. The MMR ranking method with a few existing novelty models is presented as the solution for the problem. Then, the intertemporal choice browsing model is proposed to be used to improve the MMR ranking method.

Based on the findings, the conventional ranking system is indeed less effective for post retrieval browsing. It shows that the dependent relevance assumption captured by the novelty models could improve the effectiveness of the system. In addition, the combination of novelty and relevance could further improve the performance of the system. However, the best configured intertemporal choice browsing model is unable to further improve the performance of the system.

The performance of the intertemporal choice browsing model in this evaluation is somehow similar to its performance of the model in previous evaluation, where in both cases, the model is unable to produce an optimal recommendation. Moreover, the

model has been tested on two different applications with two different evaluation strategies. It is very important that further investigation on this model is conducted based on user study and not based on simulated study. User study might be able to provide a better overview on the actually usability of the intertemporal choice browsing model for IR application.

7 Conclusion and Future Work

7.1 Conclusion

This thesis investigates the effectiveness of the intertemporal choice model in modelling browsing behaviour of the user. It is based on the assumption that the problem of browsing is an intertemporal choice problem. Based on the model, the *utility* of browsing can be estimated to reflect the usefulness of a browsing strategy to the user. Therefore, a browsing recommendation model can be developed based on the intertemporal choice model to identify a good browsing strategy for the user.

The intertemporal choice model has been extensively studied for over 70 years (Loewenstein and Prelec, 1992). The effectiveness of the model in modelling decision behaviour of individual has been realised in other domains (Frederick et al., 2002). Therefore, the incorporation of the intertemporal choice model in an information retrieval (IR) system could improve the effectiveness of browsing for the system.

This thesis also investigates the effectiveness of using the implicit relevance feedback (RF) system as a retrieval system for mobile devices. It argues that the implicit RF system could overcome the problem of small screen size and limited interaction capability of the mobile devices in searching for information. Moreover, the implicit RF system provides a platform to evaluate the effectiveness of the intertemporal choice browsing model proposed in this thesis.

This thesis was organized as follows. In Chapter 1, an introduction to the main problem investigated in this thesis is provided, which is the problem of modelling browsing behaviour of the user. In particular, this thesis investigates the effectiveness of modelling browsing behaviour by using the intertemporal choice model. This thesis argues that the *incorporation of an intertemporal choice model into an IR system can improve the effectiveness of browsing for user using the system.*

Then, a detailed discussion on the research related to the problem investigated in this thesis is provided in Chapter 2. It includes various issues such as the relevance feedback system, the problem of modelling browsing behaviour and most importantly a detailed review of the intertemporal choice model.

After that, a detailed discussion on the behavioural model for intertemporal choice (Loewenstein and Prelec, 1992) adopted in this thesis is provided in Chapter 3. The issues related to the application of the model for modelling browsing behaviour are also described. One of the issues is the technique to estimate the value of a document. Two dependent relevance models are proposed, the collective value and the differential value for a document. A collective value of a document is measured by combining its value with the value of the previously read documents in a browsing session. Meanwhile, the differential value of a document is the value measured against the documents previously read, such as based on the novel information it contains. Moreover, the intertemporal choice model is amended to model uncertainty in the estimation of value for a document.

Next, the effectiveness of an implicit RF system is investigated in Chapter 4. The main reason for the investigation is to provide a platform to evaluate the intertemporal choice browsing model proposed in Chapter 3. It seems that the effectiveness of the implicit RF system depends on the effectiveness of the intertemporal choice browsing model in choosing a good browsing strategy for the user. In addition, a number of implicit RF models and display strategies are also investigated to optimise the performance of the system.

Then, in Chapter 5 the effectiveness of the intertemporal choice browsing model is evaluated for browsing in the implicit RF system investigated in Chapter 4. The role of the model is to choose one browsing path from all possible browsing paths of the

system. In order for the implicit RF system to be effective, the intertemporal choice browsing model should suggest a good browsing path for the user. Various implementation issues are also investigated such as parameter tuning for the discount function and modelling uncertainty for the value of the documents.

Finally, Chapter 6 provides another platform to evaluate the effectiveness of the intertemporal choice browsing model. In this chapter, the effectiveness of the model is investigated for the post retrieval browsing scenario in the context of the subtopic relevance retrieval application (Zhai et al., 2003). The goal of the intertemporal choice browsing model is to produce an optimal ranking for sequential assessment of the top retrieved documents.

The main findings of this thesis are summarised in this chapter. Then, a general conclusion of this thesis is provided. After that, the contributions of this thesis to IR research are presented. Finally, some possible improvements of this work are suggested for future investigation.

7.1.1 An implicit RF system for mobile devices

In this thesis, the problem of implementing an effective IR system for mobile devices is investigated. The use of an implicit relevance feedback (RF) system for mobile devices was first proposed in (Vinay et al., 2005). Nevertheless, the use of the ostensive model (Campbell and van Rijsbergen, 1996) as implicit RF model for such an application is first investigated in this thesis. The effectiveness of the ostensive model has been investigated in a similar application, the Ostensive Browser, for image retrieval (Campbell, 2000). In addition, this thesis evaluated the effectiveness of the implicit RF system on reasonably large test collections with proper queries.

The effectiveness of the implicit RF system is measured by the mean average precision (MAP), of all possible browsing paths that can be generated by the system for a given query. The average MAP score of the system indicates its average performance. It is compared to the MAP score of the browsing path generated by the baseline system, in which the top ranked documents are displayed in batches to the user. Based on the evaluation, the average performance of the implicit RF system is not necessarily more effective than the performance of the baseline.

Nevertheless, the implicit RF system has a potential to improve the effectiveness of the baseline. It has been shown that there exist some of the possible browsing paths generated by the system that could potentially lead to a better MAP score as compared to the baseline, provided that the user chooses the right documents while browsing. This is problematic, since the user is usually unfamiliar with the system and he/she does not have enough knowledge to make an optimal decision. Hence, there is no guarantee that the user will benefit from the implicit RF system. This problem might be overcome by incorporating an effective browsing recommendation system into the application to suggest an ideal browsing strategy for the user.

In addition, the evaluation conducted in this thesis also suggests that the increasing profile of the ostensive model is better than the constant and the decreasing profile (Campbell, 2000). This finding contradicts the claims made in (Campbell, 2000) and (Hirashima et al., 1998), which suggested that the decreasing profile is more suitable. The increasing profile suggested that a recently viewed document should be weighted less than the older or past document. A similar profile is used in (White et al., 2004) but no comparison among the three profiles was conducted. In the context of the evaluation conducted in this thesis, the increasing profile is consistently better than the other profiles.

To summarise, the implicit RF system could be a solution for implementing IR systems on mobile devices. However, an effective browsing recommendation model is required to ensure that the user will fully benefit from using the system. In the context of this thesis, the implicit RF system for mobile devices provides an opportunity to evaluate the effectiveness of modelling browsing behaviour of users by using the intertemporal choice model, which will be discussed in the next section.

7.1.2 An effective recommendation model for the implicit RF system

As part of the solution for the implicit RF system on mobile devices discussed earlier, this thesis investigated the effectiveness of modelling browsing behaviour of users based on the intertemporal choice model in an attempt to develop an effective browsing recommendation model for the implicit RF system. It is based on the assumption that the preference of the users for a sequence of documents can be modelled by the

intertemporal choice model such that an ideal browsing path (sequence of documents) can be identified easily.

The effectiveness of the browsing recommendation model is evaluated based on the browsing path suggested. It is compared against two common browsing strategies, random browsing and browsing by simply choosing the top documents displayed on the screen. In addition, the MAP score of the recommended browsing path is compared to the average MAP score of the implicit RF system in order to discover its benefit against the average browsing without the benefit of recommendations. Ultimately, the effectiveness of the recommended browsing path is compared to the baseline in order to observe the benefit of using the implicit RF system.

It has been shown that the browsing recommendation model is able to suggest a browsing strategy or browsing path that is as good as the average browsing. In addition, the recommended browsing path is more effective than the browsing path generated by random browsing or by simply choosing the top documents. In the context of this evaluation, the benefit of the browsing recommendation model in the implicit RF system is truly realised.

However, the main role of the browsing recommendation model is to suggest a browsing path that is better than the browsing path generated by the baseline system. If this condition is satisfied, the real benefit of the implicit RF system will be realised. However, the browsing recommendation model is not necessarily able to produce a browsing path that is significantly more effective than the browsing path of the baseline, although in some cases a better browsing path is recommended. Moreover, in the context of the evaluation conducted, the recommended browsing path is as good as the browsing path of the baseline.

In this thesis, the cost of browsing is measured by the expected search length (ESL) for the browsing path. The cost is associated with unnecessarily downloading non-relevant documents. Therefore, the aim is to minimise the expected number of non-relevant documents to be assessed before a certain number of relevant documents is found. For the evaluation, the ESL is measured based on finding the first relevant document. Based on this measure, the browsing recommendation model is less effective than the baseline.

In this thesis, the intertemporal choice model is amended to model uncertainty in the estimation of the value for a document. Unlike the monetary system, where the value of money is certain, the value estimated for a document may not be certain. For instance, the value of a document could depend on the context of the user's query. Therefore, a document could have more than one value. In such cases, the model suggests that an expected value for the document should be used. Based on the results of the experiment, such a technique may improve the performance of the recommendation model in some cases. In particular, such an improvement may depend on the test collection used.

As a result, the potential benefit of the intertemporal choice model may not be realised in the context of this evaluation. One of the problems is that the model uses the retrieval status value (RSV) of a document assigned by the IR system as the exact value for the document. Therefore, it is impossible to know whether the poor performance is due to the model itself or due to the RSV value of the documents. In order to investigate this scenario, the RSV value of the document is replaced by the relevance value for the document given by the test collection, which is the QRel value. As a result, a marginal improvement of the browsing recommendation model performance can be observed. However, with such a perfect setting, the recommendation model still fails to produce the best recommendation of browsing path for the system, even though the performance is much better than the baseline.

As a summary, the proposed recommendation model based on the intertemporal choice model may improve the performance of the implicit RF system provided that the model is set to its optimal setting. However, there is a high potential that the model will fail to find the optimal browsing path for the user of the implicit RF system.

7.1.3 An effective ranking for post retrieval browsing

Ranking the retrieved documents in decreasing order of their relevance values has been suggested as the optimal ranking strategy (Robertson, 1977, Gordon and Lenk, 1991). It maximised the expectation of finding relevant documents. However, assessing the documents by following their ranks may not be an optimal strategy since the documents are not assessed independently (Gordon and Lenk, 1992). Therefore, an alternative ranking method is required for post-retrieval browsing.

The problem of post-retrieval browsing is investigated in the context of the subtopic relevance retrieval application (Zhai et al., 2003), where a topic consists of a set of subtopics and a document can be relevant to none and up to all subtopics of a given topic. The goal of the ranking of the subtopic relevance retrieval application, assuming the user is browsing through the ranking, is to cover many subtopics as soon as possible.

Based on the experiment conducted, it has been shown in the evaluation that the PRP-based ranking is not optimal for post-retrieval browsing since the documents are not assessed independently. In addition, ranking the documents based on a differential value, such as novelty or the Maximal Marginal Relevance (MMR) score, is better for post-retrieval browsing.

Next, the use of the intertemporal choice model in this context was evaluated. The ranking produced by this model was compared against ranking by the RSV score, by novelty and combinations of RSV score and novelty score (i.e. MMR). The intertemporal choice model could be shown to produce a better ranking for post-retrieval browsing in some circumstances, provided that the model parameters are carefully tuned.

In addition, not all novelty models can produce a better ranking for post retrieval browsing. However, measuring novelty based on the number of new words in a document is shown to be effective in this evaluation.

7.1.4 General conclusion

This thesis argues that the *incorporation of an intertemporal choice model into an IR system can improve the effectiveness of browsing for user using the system.* The effectiveness of the intertemporal choice model in modelling the decision behaviour of individuals has been established in other domains based on many empirical studies (Frederick et al., 2002). Therefore, it is hypothesised that an effective browsing recommendation can be achieved by incorporating the intertemporal choice model into an IR system.

In this thesis, the browsing recommendation problem is investigated in the context of browsing on an implicit RF system for mobile devices and post-retrieval browsing. Experiments showed that the proposed model has a potential to improve the effectiveness of browsing for an IR system. However, the effectiveness of the proposed

model depends on the quality of the model to estimate the value of documents and also its parameter setting. These limitations need to be optimised in order to guarantee the effectiveness of the proposed recommendation model. Moreover, based on the experiments conducted, there is a chance that the model will not be able to recommend the best browsing path for the user.

Further investigation based on user study would be a reasonable alternative before a conclusion can be made on the applicability of the intertemporal choice browsing model for solving IR problem. In particular, the ability of the intertemporal choice model to model browsing behaviour of the user can only be verified by user study.

7.2 Contributions of this thesis

The main contribution of this thesis is the application of the intertemporal choice model to solve an IR problem. In particular, this thesis proposed a browsing recommendation model to improve the effectiveness of the user in browsing, by adopting a behavioural model based on intertemporal choice. Based on the evaluation conducted in this thesis, the proposed model has a potential to improve the effectiveness of browsing in an IR system.

In particular, the effectiveness of an implicit RF system for mobile devices is investigated. The system is evaluated on large test collections with proper set of queries. The use of the ostensive model as implicit RF model for the system is also investigated. Furthermore, a browsing recommendation model is proposed to ensure a better performance for the system by suggesting a good browsing strategy for the user.

Additionally, the effectiveness of a conventional ranking system is investigated for post-retrieval browsing. The investigation showed that the conventional ranking system is not effective for post-retrieval browsing. Alternative ranking systems based on modelling novelty are more effective for post-retrieval browsing. In addition, the effectiveness of a number of novelty models is also studied.

7.3 Future work

Due to the limitations of the intertemporal choice model based on the experiments conducted in this thesis, two new areas of research can be introduced. First, the effectiveness of the model depends on the quality of the model to estimate the value of a document. It has been shown that a method that can accurately estimate the value of a

document will improve the effectiveness of the model. However, the problem of assigning a good value to a document is an open research question in IR. For a start, we can investigate the effectiveness of all available value estimation models in IR in dealing with this particular problem. In addition, a carefully designed user study similar to the design used in the economic domain can be conducted to understand how the value of a document is perceived by the user. Such an approach has not been investigated in IR.

Second, the effectiveness of the intertemporal choice model depends on the parameter setting for the discount function. The evaluation in this thesis indicated that, the parameters play an important role in the performance of the intertemporal choice model. It will be interesting to discover the optimal setting for these parameters. In addition, the value of the parameter defines the discount rate of a discount function. Therefore, a model to predict the optimal value for the parameter is required in order to ensure the usefulness of the intertemporal choice model. Furthermore, a powerful prediction model could make the intertemporal choice model more adaptive towards the problem that is being modelled. For instance, an optimal value for the parameter may depend on the query or the collection. In such a case, the prediction model could make the intertemporal choice model more adaptive to the environment of the decision problem.

Finally, this thesis focuses on the application of the model to an existing problem in IR. The model is tuned such that an optimal performance can be achieved. However, the development of the model in the context of IR is not investigated. For instance, the ideal discount function or discount rates cannot be realised without proper user studies. Therefore, further work to this end should focus also on the development aspect of the model in the context of IR, based on studying real user behaviour in IR situations.

Bibliography

Agostini, A. and Avesani, P. (2003) Advertising games for Web services, In *Proceedings of the Eleventh International Conference on Cooperative Information Systems (CoopIS-03)*.93-109

Agostini, A. and Avesani, P. (2004) On the discovery of the semantic context of queries by game-playing, In *Proceedings of the Sixth International Conference on Flexible Query Answering Systems (FQAS-04)*,

Albers, M. J. and Kim, L. (2000) User web browsing characteristics using palm handhelds for information retrieval, In *Proceedings of IEEE professional communication society international professional communication conference*.125-135

Allan, J., Wade, C. and Bolivar, A. (2003) Retrieval and Novelty Detection at the Sentence Level, In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*.314 - 321

Armstrong, R., Freitag, D., Joachims, T. and Mitchell, T. (1995) WebWatcher: A Learning Apprentice for the World Wide Web, In *AAAI Spring Symposium on Information Gathering*,

Avesani, P. and Agostini, A. (2003). A peer-to-peer advertising game. In *First International Conference on Service Oriented Computing (ICSOC-03)* (Eds, Orlowksa, M., Papazoglou, M., Weerawarana, S. and Yang, J.), Berlin Heidelberg: Springer-Verlag LNCS 2910

Azman, A. and Ounis, I. (2004) Discovery of aggregate usage profiles based on clustering information needs, In *Proceedings of the 27th annual international conference on Research and development in information retrieval*.470 - 471

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

Beaulieu, M. (1998). Experiments on the interfaces to support query expansion. *Journal of Documentation,* **53 (1)**.8-19.

Belkin, N. J., Oddy, R. N. and Brooks, H. M. (1982). ASK for information retrieval: Part I - background and theory. *Journal of Documentation,* **38 (2)**.61-71.

Benzion, U., Rapoport, A. and Yagil, J. (1989). Discount rates inferred from decisions: An experimental study. *Management Science,* **35 (3)**.270-284.

Bollmann, P. and Raghavan, V. V. (1988) A utility-theoretic analysis of expected search length, In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*.245-256

Borlund, P. (2003). The concept of relevance in Information Retrieval. *Journal of the American Society for Information Science and Technology,* **54 (10)**.913-925.

Brin, S. and Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks,* **30 (1-7)**.107-117.

Buyukkokten, O., Garcia-Molina, H. and Paepcke, A. (2001) Seeing the whole in parts: text summarization for web browsing on handheld devices, In *Proceedings of the 10th international conference on World Wide Web*.652 - 662

Buyukkokten, O., Garcia-Molina, H., Paepcke, A. and Winograd, T. (2000) Power browser: efficient Web browsing for PDAs, In *Proceedings of the SIGCHI conference on Human factors in computing systems*.430-437

Buyukkokten, O., Kaljuvee, O., Garcia-Molina, H., Paepcke, A. and Winograd, T. (2002). Efficient Web Browsing on Handheld Devices Using Page and Form Summarization. *ACM Transactions on Information Systems,* **20 (1)**.82-115.

Cairns, J. and van der Pol, M. (2000). Valuing further private and social benefits: The discounted utility model versus hyperbolic discounting models. *Journal of Economic Psychology,* **21**.191-205.

Campbell, I. (2000). Interactive evaluation of the Ostensive Model using a new test collection of images with multiple relevance assessment. *Information Retrieval,* **2 (1)**.87-114.

Campbell, I. and van Rijsbergen, C. J. (1996). The ostensive model of developing information needs. In *Proceedings of {COLIS}-96, 2nd International Conference on Conceptions of Library Science*, 251-268.

Carbonell, J. and Goldstein, J. (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries, In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*.335 - 336

Chang, Y., Ounis, I. and Kim, M. (2006). Query reformulation using automatically generated query concepts from a document space. *Information Processing & Management,* **42 (2)**.453-468.

Chen, M.-S., Park, J. S. and Yu, P. S. (1998). Efficient Data Mining for Path Traversal Patterns. *IEEE Transactions on Knowledge and Data Engineering,* **10 (2)**.209-221.

Chen, Y., Ma, W.-Y. and Zhang, H.-J. (2003) Detecting web page structure for adaptive viewing on small form factor devices, In *Proceedings of the 12th international conference on World Wide Web*.225-233

Chi, E. H., Pirolli, P., Chen, K. and Pitkow, J. E. (2001) Using information scent to model user information needs and actions and the Web, In *CHI 2001*.490-497

Chi, E. H., Pirolli, P. and Pitkow, J. (2000) The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a Web site, In *The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a Web site*.161-168

Cleverdon, C. W. (1997). The Cranfield test on index language devices. In *Readings in Information Retrieval* (Eds, Sparck Jones, K. and Willet, P.), 47-59. San Francisco: Morgan Kaufmann Publishers

Cooper, W. S. (1968). Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems. *American Documentation,* **19 (1)**.30-41.

Cooper, W. S. and Maron, M. E. (1978). Foundations of Probabilistic and Utility-Theoretic Indexing. *Journal of the ACM (JACM),* **25 (1)**.67-80.

Croft, W. B. and Thompson, R. H. (1987). I3R: A new approach to the design of the document retrieval systems. *Journal of The American Society for Information Science,* **38 (6)**.389-404.

Croft, W. B. and van Rijsbergen, C. J. (1976). An Evaluation of Goffman's Indirect Retrieval Method. *Information Processing & Management,* **12**.327-331.

de Bruijn, O., Spence, R. and Chong, M. Y. (2002). RSVP Browser: Web Browsing on Small Screen Devices. *Personal and Ubiquitous Computing,* **6 (4)**.

Dunham, M. H. (2003). Web Mining. In *Data Mining: Introductory and Advanced Topics*, 195-220. Prentice Hall

Ellis, D. (1989). A behavioral approach to information retrieval system design. *Journal of Documentation,* **45 (3)**.171-212.

Frederick, S., Loewenstein, G. and O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature,* **40 (2)**.351-401.

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

Gabrilovich, E., Dumais, S. and Horvitz, E. (2004) Newsjunkie: providing personalized newsfeeds via analysis of information novelty, In *Proceedings of the 13th international conference on World Wide Web*.482 - 490

Gery, M. (2002). Non-linear reading for a structured web indexation. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 379-380. Tampere, Finland: ACM Press

Goffman, W. (1969). An Indirect Method of Information Retrieval. *Information Storage and Retrieval,* **4**.361-373.

Golovchinsky, G. (1997) What the query told the link: The integration of the hypertext and information retrieval, In *Proceedings of the 8th ACM Conference on Hypertext*,

Gordon, M. D. and Lenk, P. (1991). A Utility Theoretic Examination of the Probability Ranking Principle in Information Retrieval. *Journal of The American Society for Information Science,* **42 (10)**.703-714.

Gordon, M. D. and Lenk, P. (1992). When is the Probability Ranking Principle Suboptimal. *Journal of The American Society for Information Science,* **43 (1)**.1-14.

Green, L. and Myerson, J. (1996). Exponential Versus Hyperbolic Discounting of Delayed Outcomes: Risk and Waiting Time. *American Zoologist,* **36 (4)**.496-505.

Harman, D. (1992) Relevance feedback revisited, In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*.1-10

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

Harvey, C. M. (1986). Value functions for infinite-period planning. *Management Science,* **32 (9)**.1123-1139.

Hearst, M. A. and Pedersen, J. O. (1996) Reexamining the cluster hypothesis: scatter/gather on retrieval results, In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*.76-84

Hirashima, T., Matsuda, N., Nomoto, T. and Toyoda, J. i. (1998). Context-sensitive filtering for browsing in hypertext. In *Proceedings of the 3rd international conference on Intelligent user interfaces*, 119-126. San Francisco, California, United States: ACM Press

Huberman, B. A., Pirolli, P., Pitkow, J. E. and Lukose, R. M. (1998). Strong Regularities in World Wide Web Surfing. *Science,* **280**.95-97.

Joachims, T., Freitag, D. and Mitchell, T. M. (1997). Web Watcher: A Tour Guide for the World Wide Web. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97)*, 770-777.

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K. and Buchanan, G. (1999). Improving Web interaction on small displays. *Computer Networks: The International Journal of Computer and Telecommunications Networking,* **31 (11-16)**.1129-1137.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica,* **47 (2)**.263-292.

Kelly, D. and Teevan, J. (2003). Implicit Feedback for Inferring User Preference: A Bibliography. *ACM SIGIR Forum,* **37 (2)**.18-28.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM),* **46 (5)**.604-632.

Koenemann, J. and Belkin, N. J. (1996) A case for interaction: a study of the interactive information retrieval behavior and effectiveness, In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*.205-212

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R. and Riedl, J. (1997). GroupLens: Applying collaborative filtering of news. *Communications of the ACM,* **40 (3)**.77-87.

Lai, A. M., Nieh, J., Bohra, B., Nandikonda, V., Surana, A. P. and Varshneya, S. (2004) Improving web browsing performance on wireless PDAs using thin-client computing, In *Proceedings of the 13th international conference on World Wide Web*.143-154

Lazaro, A., Barberan, R. and Rubio, E. (2002). The discounted utility model and social preferences: Some alternative formulations to conventional discounting. *Journal of Economic Psychology,* **23**.317-337.

Leuski, A. (2000) Relevance and Reinforcement in Interactive Browsing, In *Proceedings of the ninth international conference on Information and knowledge management*.119-126

Lieberman, H. (1995) Letizia: An Agent That Assists Web Browsing, In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence ({IJCAI}-95)*.924-929

Liu, H., Xie, X., Ma, W.-Y. and Zhang, H.-J. (2003) Automatic browsing of large pictures on mobile devices, In *Proceedings of the eleventh ACM international conference on Multimedia,*

Loewenstein, G. and Prelec, D. (1991). Negative time preference. *The American Economic Review,* **81 (2)**.347-352.

Loewenstein, G. and Prelec, D. (1992). Anomalies in Intertemporal Choice: Evidence and Interpretation. *The Quarterly Journal of Economics,* **107 (2)**.573-597.

Loewenstein, G. F. and Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review,* **100 (1)**.91-108.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development,* **2**.159-165.

Maron, M. E. and Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM),* 7.216-244.

Mobasher, B., Dai, H., Luo, T. and Nakagawa, M. (2001) Effective Personalization Based on Association Rule Discovery from Web Usage Data, In *The 3rd ACM Workshop on Web Information and Data Management (WIDM01)*.9-15

Mobasher, B., Dai, H., Luo, T. and Nakagawa, M. (2002). Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery,* **6 (1)**.61-82.

Morita, M. and Shinoda, Y. (1994) Information filtering based on user behavior analysis and best match retrieval, In *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*.272-281

Myerson, J. and Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior,* **64 (3)**.263-276.

Oddy, R. N. (1977). Information retrieval through man-machine dialogue. *Journal of Documentation,* **33 (1)**.1-14.

Olston, C. and Chi, E. H. (2003). ScentTrails: Integrating Browsing and Searching on the Web. *ACM Transactions on Computer-Human Interaction,* **10 (3)**.177-197.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C. and Johnson, D. (2005) Terrier Information Retrieval Platform, In *In Proceedings of the 27th European Conference on Information Retrieval (ECIR 2005)*.21-23

Perkowitz, M. and Etzioni, O. (2000). Towards adaptive Web sites: Conceptual framework and case study. *Artificial Intelligent,* **118**.245-275.

Pirolli, P. (1998) Exploring browser design trade-offs using a dynamical model of optimal information foraging, In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'98)*.33-40

Pirolli, P. and Card, S. K. (1995). Information foraging in information access environments. In *Proceedings of the CHI'95, ACM Conference on Human Factors in Software*, 51-58.

Pirolli, P. and Card, S. K. (1999). Information Foraging. *Psychological Review,* **106 (4)**.643-675.

Pirolli, P., Schank, P., Hearst, M. and Diehl, C. (1996). Scatter/Gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI '96,*

Ponte, J. M. and Croft, W. B. (1998) A language modeling approach to information retrieval, In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,*

Porter, M. F. (1980). An algorithm for suffix stripping. *Program,* **14 (3)**.130-137.

Rachlin, H., Raineri, A. and Cross, D. (1991). Subjective probability and delay. *Journal of the Experimental Analysis of Behavior,* **55 (2)**.233-244.

Rafaeli, S. and Raban, D. R. (2003). Experimental Investigation of the Subjective Value of Information in Trading. *Journal of the Association for Information System,* **4**.119-139.

Read, D. (2004). Intertemporal choice. In *Blackwell Handbook of Judgement and Decision Making* (Eds, Koehler, D. and Harvey, N.), 424-443. Oxford: Blackwell

Robertson, S. E. (1977). The probability ranking principle in Information Retrieval. *Journal of Documentation,* **33**.294-304.

Robertson, S. E. and Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of The American Society for Information Science,* **27 (3)**.129-146.

Rodden, K., Milic-Frayling, N., Sommerer, R. and Blackwell, A. (2003) Effective Web Searching on Mobile Devices, In *Proceedings of the 17th Annual Conference on Human-Computer Interaction*.281-296

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

Salton, G. and Buckley, C. (1990). Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science,* **41 (4)**.288-297.

Salton, G., Wong, A. and Yang, C. S. (1997). A vector space model for automatic indexing. In *Readings in Information Retrieval* (Eds, Sparck Jones, K. and Willet, P.), 273-280. San Francisco: Morgan Kaufmann Publishers

Salton, G. and Wu, H. (1980) A term weighting model based on utility theory, In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*.9-22

Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies,* **4 (2)**.155-161.

Seo, Y.-W. and Zhang, B.-T. (2000) Learning user's preferences by analysing web browsing behaviors, In *Proceedings of the 4th ACM International Conference on Autonomous Agents*.381-387

Shen, X. and Zhai, C. (2005) Active Feedback in Ad Hoc Information Retrieval, In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*.59 - 66

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum,* **33 (1)**.6-12.

Smucker, M. D. and Allan, J. (2006) Find-similar: similarity browsing as a search tool, In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*.461-468

Spink, A., Greisdorf, H. and Bateman, J. (1998). From highly relevant to not relevant: examining different regions of relevance. *Information Processing & Management,* **34 (5)**.599-621.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent Systems,* **16 (5)**.16-22.

Sweeney, S. and Crestani, F. (2006). Effective search results summary size and device screen size: Is there a relationship? *Information Processing & Management,* **42 (4)**.1056-1074.

Tao, T. and Zhai, C. (2006) Regularized estimation of mixture models for robust pseudo-relevance feedback, In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*.162 - 169

Urban, J., Jose, J. M. and van Rijsbergen, C. J. (2003) An Adaptive Approach Towards Content-based Image Retrieval, In *Proceeding of the Third International Workshop on Content-Based Multimedia Indexing (CBMI'03)*,

van der Pol, M. M. and Cairns, J. A. (2000). Negative and zero time preference for health. *Health Economics,* **9**.171-175.

Van Rijsbergen, C. J. (1979) *Information Retrieval,* Butterworth & Co Ltd., London.

Van Rijsbergen, C. J. (1986a) A new theoretical framework for information retrieval, In *Proceedings of the 10th International ACM SIGIR Conference on Research and Development in Information Retrieval*.194-200

Van Rijsbergen, C. J. (1986b). A non-classical logic for information retrieval. *The Computer Journal,* **29 (6)**.481-485.

Vinay, V., Cox, I. J., Milic-Frayling, N. and Wood, K. (2005) Evaluating Relevance Feedback Algorithms for Searching on Small Displays, In *Proceedings of the European Conference on Information Retrieval (ECIR) 2005*.185-199

von Neumann, J. and Morgenstern, O. (1944) *Theory of Games and Economic Behavior,* Princeton University Press.

Wexelblat, A. and Maes, P. (1997) Footprints: History-rich web browsing, In *Computer-Assisted Information Retrieval (RIAO)*.75-84

Wheeldon, R. and Levene, M. (2003) The best trail algorithm for assisted navigation of Web sites, In *Proceedings of the First Latin American Web Congress (LA-WEB 2003)*.166-178

White, R. W., Jose, J. M., van Rijsbergen, C. J. and Ruthven, I. (2004) A Simulated Study of Implicit Feedback Models, In *Proceedings of the 26th Annual European Conference on Information Retrieval (ECIR 2004)*.311-325

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

Zhai, C., Cohen, W. W. and Lafferty, J. (2003) Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval, In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*.10 - 17

Zhai, C. and Lafferty, J. (2006). A risk minimization framework for information retrieval. *Information Processing & Management,* **42 (1)**.31-55.

Zhang, H.-P., Xu, H.-B., Bai, S., Wang, B. and Cheng, X.-Q. (2004) Experiments in TREC 2004 Novelty Track at CAS-ICT, In *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004)*,

Zhang, Y., Callan, J. and Minka, T. (2002) Novelty and Redundancy Detection in Adaptive Filtering, In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*.81 - 88

Zhu, T., Greiner, R. and Häubl, G. (2003) Predicting Web Information Content, In *Proceedings of the Workshop on Intelligent Techniques for Web Personalization (ITWP'03)*,

Glossary

IR	Information Retrieval
RF	Relevance Feedback
IIR	Interactive Information Retrieval
MMR	Maximal Marginal Relevance
TREC	Text Retrieval Conference
RSV	Retrieval status value
PRP	Probability Ranking Principle
DU	Discounted Utility Model
TopK	Display strategy that selects the top-*k* documents
GappedTopK	Display strategy that selects the top-*k* documents with a certain gap introduced between the ranking
DT	Decision tree
Q	Query
MAP	Mean average precision measure
ESL	Expected search length measure
OMDec	The ostensive model with decreasing profile
OMCon	The ostensive model with constant profile
OMInc	The ostensive model with increasing profile
DN1	Novelty model: average cosine distance
DN2	Novelty model: minimum cosine distance
DN3	Novelty model: average Kullback-Leibler similarity
DN4	Novelty model: maximum Kullback-Leibler similarity
SN1	Novelty model: new word count
SN2	Novelty model: new word count normalised by the length
SN3	Novelty model: weighted new word count
SN4	Novelty model: one minus the overlap ratio
s_recall	Subtopic recall
s_precision	Subtopic precision
SingleValue	Technique that chooses the value of the selected documents in a browsing session as the values for the browsing path
DecreasingWeight	Technique that chooses the combination values of the displayed documents with a decreasing weight to the document as its rank number increases
EqualWeight	Technique that chooses the average value of the displayed documents as the values for the browsing path
QRel	The relevance judgement data from the test collection