HEIKKI KESKUSTALO

# Towards Simulating and Evaluating User Interaction in Information Retrieval using Test Collections

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Sciences of the University of Tampere,
for public discussion in the Auditorium Pinni B 1096,
Kanslerinrinne 1, Tampere, on November 27th, 2010, at 12 o'clock.

UNIVERSITY OF TAMPERE

UNIVERSITY
OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere
Department of Information Studies and Interactive Media
Finland

Distribution
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 40 190 9800
Fax +358 3 3551 7685
taju@uta.fi
www.uta.fi/taju
http://granum.uta.fi

Cover design by
Mikko Reinikka

Acta Universitatis Tamperensis 1563
ISBN 978-951-44-8262-5 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1012
ISBN 978-951-44-8263-2 (pdf)
ISSN 1456-954X
http://acta.uta.fi

Tampereen Yliopistopaino Oy – Juvenes Print
Tampere 2010

# Acknowledgements

First of all, I wish to express my gratitude to my supervisors Kalervo Järvelin and Ari Pirkola. This research was mainly conducted during the period when Kalle was the Academy Professor of the Academy of Finland. I have enjoyed Kalle’s intellectual openness, uncompromising professionalism, independent thinking and joyful creativity for many years. Ari’s kind guidance and expert comments have been essential in my personal research journey.  I feel lucky that I have had Kalle and Ari as my supervisors.

Next I would like to thank my co-authors Jaana Kekäläinen, Tarun Sharma and Marianne Lykke for good cooperation and team spirit. I thank all my colleagues in FIRE (Finnish Information Retrieval Expert Group) who commented my thesis, especially Kai Halttunen, Per Ahlgren, Marjo Markkula, Eija Airio, Paavo Arvola, Feza Baskaya, Turid Hedlund, Aki Loponen, and Sanna Kumpulainen.  Especially, I wish to thank Eero Sormunen and Pertti Vakkari for their insightful comments. Eero’s and Pertti’s studies in the areas of graded relevance and user interaction have an important role in the present study.

I thank the pre-reviewers of my thesis, Professor Ian Ruthven, University of Strathclyde, UK, and Dr. Falk Scholer, RMIT University, Australia, for their comments. I would also like to thank very warmly Tuula Väisänen and Sanna Yli-Salomäki for their expertise and kind help in various administrational matters.

For the financial support I express my gratitude to the Academy of Finland and Tampere Graduate School in Information Science (TISE). I thank Dr. Pertti Koivisto from TISE for providing a very good research atmosphere.

My heartfelt thanks are due to my family and friends. I thank my parents, Eeli and Terttu, for encouraging me in my studies, and my sister Tuija and her family. Last, I express my loving gratitude to my wife Tarja for her continuous support, and to our son Henrikki, 5 years, for his patience and impatience.

Tampere 15.10.2010

Heikki Keskustalo

# Abstract

This thesis aims at extending traditional test collection-based evaluation (TCE) experiments of information retrieval (IR) towards real life usage while remaining within the bounds of TCE. In traditional TCE there are no interactive search processes, nor explicit assumptions of users. Instead, batch-mode retrieval experiments are assumed entailing one query per topic, well-defined and relatively verbose requests, and binary relevance judgments. In real life, on the contrary, interaction is vital. The users interact with IR systems by using a trial-and-error process trying out multiple query candidates; they vary their browsing effort, and may require only few, highly relevant documents. Importantly, users as well as searching situations may differ from each other in many ways. The individual studies of the thesis focus on query-based interaction using simulations.

Two different types of interaction simulations are performed: relevance feedback (RF) and session strategy (SS) simulations. In both cases more than one query per query session is used. In RF simulations the initial query is modified by adding feedback terms gathered automatically from relevant documents observed by the simulated user. The interaction decisions include the eagerness of the user to browse the list of retrieved documents; the effort to give document-level relevance feedback, and the relevance threshold to accept a document as feedback. These attributes are justified based on literature. In SS simulations direct query reformulations are performed based on prototypical query modifications. We also introduce the concept of negative higher-order relevance, and discuss evaluation issues when interaction and graded relevance judgments are brought to the setting.

Our main experimental results suggest that mixed-quality RF is more effective than an attempt to use solely highly relevant feedback, and that sequences of very short queries are surprisingly effective in finding relevant documents. Because interaction is an essential property of system usage in real life, we suggest that in the future test collection-based IR research should not continue excluding interaction but instead bring interaction simulations into the research forefront.

# Table of Contents

# Original publications

This thesis consists of a summary and the following original research publications:

Keskustalo, H., Järvelin, K., Pirkola, A. (2006) The Effects of Relevance Feedback Quality and Quantity in Interactive Relevance Feedback: A Simulation Based on User Modeling. In: Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsirika, T., Yavlinsky, A., eds., *Proceedings of the 28th European Conference on IR Research (ECIR'06)*, Lecture Notes in Computer Science, 3936, Springer-Verlag, Heidelberg, pp. 191-204.

Keskustalo, H., Järvelin, K., Pirkola, A. (2008a) Evaluating the Effectiveness of Relevance Feedback Based on a User Simulation Model: Effects of a User Scenario on Cumulated Gain Value. *Information Retrieval*, 11 (3), pp. 209-228.

Keskustalo, H., Järvelin, K., Pirkola, A., Kekäläinen, J. (2008b) Intuition-Supporting Visualization of User's Performance Based on Explicit Negative Higher-Order Relevance. In: Myaeng, S-H., Oard, D.W., Sebastiani, F., Chua, T-S., Leong, M-K., eds., *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, Singapore, pp. 675-681.

Keskustalo, H., Järvelin, K., Pirkola, A., Sharma, T., Lykke, M. (2009) Test Collection-Based IR Evaluation Needs Extension Toward Sessions – A Case of Extremely Short Queries. In: Lee, H. G., Song, D., Lin, C-Y., Aizawa, A. N., Kuriyama, K., Yoshioka, M., Sakai, T., eds., *Information Retrieval Technology: The 5th Asia Information Retrieval Symposium (AIRS'09)*, Lecture Notes in Computer Science, 5839, Springer-Verlag, Heidelberg, pp. 63-74.

The publications will be referred to as Studies I-IV in the summary. Reprinted by permission of the publishers.

# 1. Introduction

The fundamental purpose of an information retrieval (IR) system is to help its user finding useful information contained in documents. In the 1960s the Cranfield experiments formed the prototype of future test collection-based IR testing to come. In it, test questions are constructed, a document database is built, and intellectual relevance decisions are acquired so that the correct solution for each retrieval task is known beforehand. (Cleverdon, 1991; Voorhees, 2007)

The traditional test collections based on the Cranfield paradigm have some well known limitations. In brief, real users are not directly involved and batch-mode experiments are performed based on one query per topic (Kekäläinen and Järvelin, 2002a). In real life, on the contrary, various kinds of users are involved and various searching situations occur. Interaction is essential in real life because the initial query may fail for many reasons. Therefore, a searcher may repeatedly launch a query followed by browsing the result, until the session ends more or less successfully or the user quits without success. We will return to these issues in more detail later in this chapter.

In this thesis test collection-based evaluation (TCE) refers to the traditional Cranfield-style information retrieval evaluation experiments performed in a controlled laboratory setting. The relevance of every document in the test collection is known with respect to each pre-defined topic. Given such a collection researchers can reuse it and compare the effectiveness of alternative IR approaches.

The main justification of the TCE model comes from the observation that character string-based matching of the topics of requests and documents can be performed successfully. Natural language words extracted from the free text of the documents can be used as search keys to define topical requests because the existence of the words in the documents correlates with fair probability to the topical content of the documents they represent (Kekäläinen and Järvelin, 2002a; Ingwersen and Järvelin, 2005). Yet character string-based matching is challenging because natural language allows both searchers and authors to select concepts and express them in many ways. Moreover, the expressions and the fragments of expressions may be ambiguous. To make the situation even more complex, individual people may have different opinions regarding *what* should be retrieved (Voorhees, 2007) and *how* to retrieve it. Finally, various data structures and algorithms are required to allow rapid searching as vast amounts of text need to be searched.

The goals of research in TCE entail gaining theoretical understanding of the basic problems of IR (representing requests and documents; matching them); developing theories and methods to deal with these problems; and developing matching methods successful in practice in serving the needs of the users (Ingwersen and Järvelin, 2005). Within the scope of a typical TCE experiment a set of topical queries is run as a batch-mode experiment; the test collection is a free-text collection, consisting of, e.g., newspaper articles; and the evaluation is based on relevance judgments made by human judges acting as relevance assessors. The size of standard ad hoc test collections has grown to millions of documents, and hundreds of topics may be used in experiments. The performance measures typically focus on the ranked list or set of documents retrieved, and the effectiveness is measured in terms of available single-query metrics averaged over the set of topics.

The following list of limitations of the TCE approach is mainly based on Kekäläinen and Järvelin (2002a), and Ingwersen and Järvelin (2005). In the *traditional test collection-based evaluation* experiments, typically,

- users are abstracted away with their varying tasks and situations (Voorhees, 2007)
- well-defined information needs are assumed (Kekäläinen and Järvelin, 2002a)
- single-query per topic is assumed (no interaction) (Bates, 1989; Belkin et al., 1993)
- relatively verbose queries are used (Jansen et al., 2000)
- relevance is topical and static (Kekäläinen and Järvelin, 2002b)
- varying degrees of relevance are collapsed into a binary scale (Kekäläinen and Järvelin, 2002b)
- low relevance threshold is used (Voorhees, 2001; Sormunen, 2002; Kekäläinen and Järvelin, 2002b; Scholer, 2009)

- evaluation is based on an averaged per-query viewpoint (Rocchio, 1971a)
- evaluation is based on the quality of the ranked list retrieved (Kekäläinen and Järvelin, 2002a)
- independency of the retrieved documents is assumed (Kekäläinen and Järvelin, 2002a)
- long lists of retrieved documents may form the basis of system evaluation

Recent studies have shown that the batch-style evaluation results may differ from the evaluation results of situations where real users are involved. Real users can successfully compensate[1] for the system performance differences observed in non-interactive batch experiments.

*Real life* information retrieval can be characterized as follows:

- different searchers, tasks, and situations exist (Spink, 1997; Stenmark, 2008)
- real information needs[2] may be ill-defined; even if they are well-defined they may be difficult to express (Belkin, 1980); the needs may change (Bates, 1989); as well as the searcher's ability to articulate them (Kuhlthau, 1991; Vakkari, 2000)
- searching is a dialogue between the user and the system rather than a single information need specification; the user may need to try out multiple queries (Swanson, 1977; Belkin, 1980; Bookstein, 1983)
- users often prefer short queries (Jansen et al., 2000; Vakkari, 2000) requiring reformulations (Ruthven, 2008) often performed by using few tactical term-level moves (Vakkari, 2002)
- relevance is dynamic and has several manifestations (Saracevic, 1996b; Cosijn and Ingwersen, 2000)
- documents may not be equally relevant (Saracevic, 1975)
- the searcher may prefer highly relevant documents[3] (Järvelin and Kekäläinen, 2000; Järvelin et al., 2008)

[1] The users may, e.g., issue more queries and read more documents (Smith and Kantor, 2008).

[2] A real information need is of personal interest and importance to the user as opposed to simulated needs defined in the context of a simulated work task (see Borlund, 2000) or "topical needs" manifested as topics of test collections.

[3] Partially relevant documents might be useful in some situations, e.g., at the formative stage of the problem solving process initiating the search (Spink et al., 1998).

- evaluation takes place in multiple query context (Jansen et al., 2000)
- not only the quality of the result is significant for the user (e.g., low query formulation effort may be important)
- documents are not independent[4] and the relevance judgment process may change the initial relevance criteria themselves (Beaulieu, 2000)
- only few top documents may be inspected by the user (Jansen et al., 2000; Ruthven, 2008); only these may actually matter for the user (Azzopardi, 2007)

Involvement of the users forms a continuum in IR experiments. In one extreme, in the traditional test collection-based experiments, users may not be modeled explicitly at all.[5] In the other extreme real users are unobtrusively observed in real world situations (e.g., Spink and Saracevic, 1998). Between these two extremes it is possible to study real people performing *simulated* tasks (e.g., Belkin et al., 1995). In the present study we select the fourth remaining route: we will focus on IR interaction in the lab, without users, by simulating user behavior (see, e.g., White et al., 2004)[6]. The concept of simulation will be discussed in Chapter 3.

The individual studies of this thesis model query-based[7] interaction, in a test collection environment, via simulations. Our goal is to expand TCE towards real life by including simulation of interaction. Two types of simulations are performed: relevance feedback (RF) and session strategy (SS) simulations. Both simulations focus on multiple-query situations where more than one query is allowed per topic during interaction.

Generally speaking, in case of our RF simulations the initial query is followed by gathering RF information by pointing out relevant documents serving as the source of feedback terms for the subsequent reformulated query (Studies I and II). The major interaction decisions modeled include patience to browse the retrieved result;

[4] E.g., information in separate documents may complement each other (Ruthven and Lalmas, 2003).

[5] Traditional batch experiments can be seen as limited (albeit implicit) user simulations. See Ahlgren (2004) as an example of an explicit user simulation, modeling users with varying levels of patience and valuations for different documents in a *non-interactive* setting.

[6] Our interaction simulations differ from the approach applied in the interactive track of TREC in which the interactive search process is performed by test persons (Hersh and Over, 2001). We simulate surface-level user behavior (relevance feedback, query reformulations, and results browsing) in the laboratory setting based on explicit assumptions on user behavior but without test persons actually performing interaction.

[7] Query-based systems utilize a request expression to retrieve information from its storage (Ruthven, 2008).

the effort to give document-level RF, and the threshold of accepting documents as relevant.

In case of SS simulations the effort of gathering RF documents is replaced by direct query reformulations (Study IV). Here we focus on various prototypical query reformulation strategies assuming an impatient user who attempts to find one (either at least a marginally, or a highly) relevant document. We will also discuss the novel concept of negative higher-order relevance (NHOR) currently gaining little attention in TCE (Study III).

We will address some of the limitations of the TCE model discussed above in our simulations. In the present thesis in Studies I-IV (i.e., SI-SIV) we will:

- model searchers explicitly (SI-SIV)
- simulate multiple-query interaction (SI, SII, SIV)
- utilize both verbose queries (typical in traditional TCE) (SI-SIII) and short queries (typical in real life) (SIV)
- utilize graded relevance during simulated interaction (SI-SII), and pay attention especially to highly relevant documents in evaluation (SI-SIV)
- discuss the issue of multiple query session evaluation (SI, SII, SIV)
- discuss evaluation beyond the positive relevance conception (SIII)
- focus specifically on top ranks during evaluation (SII-SIV)

Our simulation experiments are based on a traditional test collection. Therefore the topics used are well-defined and stable, and the relevance concept is topical and static, and we assume independence of the relevant documents.

Automated evaluation has a long history in IR (see, e.g., Cooper, 1973) but so far non-interactive testing has dominated test collection-based evaluation. Recently the need for simulated evaluation of interactive IR has gained growing attention (see Fuhr et al., 2009). A recent SIGIR workshop initiative (Azzopardi et al., 2010) proposes producing a survey of simulated evaluation of interactive IR, along with methodological guidelines and a road map for future research. Together the four studies of the present thesis seek to contribute to IR knowledge by expanding TCE methodology via explaining, justifying and demonstrating interaction simulations, and by discussing their limitations and future prospects. We shall look at the following research questions.

*Study I*

1. How effective is relevance feedback if we consider various thresholds of relevance in evaluation?
2. How is the quality and quantity of the RF related to retrieval effectiveness?
3. Can pseudo-RF successfully compete with the simulated RF?

In the first paper we develop and justify a simple interactive simulation model and perform a user simulation varying the quality and quantity of the user-given feedback. We compare the effectiveness of simulated user-RF to pseudo-relevance feedback using total performance evaluation[8] and a traditional effectiveness measure mean average precision (MAP).

*Study II*

In this paper we continue user RF simulations, but focus on rank-based evaluation (using the cumulated gain measure), and deepen the discussion on interactive evaluation. Our research questions are:

1. How should we evaluate the effectiveness of simulated user RF considering graded relevance assessments?
2. How successful are various RF strategies?

*Study III*

In this paper we will introduce the novel concept of *negative* higher-order relevance and perform traditional single-query experiments. Negative higher-order relevance refers to the negative aspects of beyond-topical relevance, e.g., dissatisfaction, frustration, and uncertainty experienced by the user. Our research questions are:

1. What is negative higher-order relevance and what is its justification?
2. How can we operationalize negative higher-order relevance?
3. What are the consequences of allowing explicit NHOR in IR evaluation?

[8] Evaluation concepts are discussed in more detail in Chapter 4.

*Study IV*

In this paper we demonstrate laboratory-based *session strategy* simulations based on query data collected from test persons. We construct prototypical short query sequences and test their effectiveness assuming impatient searchers. Our main research question is:

1. How effective are sequences of short queries combined with impatient browsing, compared to using one long query and patient browsing?

The rest of the Summary is organized as follows. Chapter 2 will discuss the limitations of the traditional test collection-based approach in more detail. Chapter 3 explains and justifies our simulations as a method for extending the traditional TBE approach. Chapter 4 discusses the evaluation problem in laboratory-based user interaction simulations. Chapter 5 summarizes the results, and Chapter 6 concludes the Summary with discussion and conclusions.

# 2. Test collection-based IR

## 2.1 The traditional test collection approach

The history of test collection-based information retrieval evaluation dates back to the end of the Second World War. In that time, vast amounts of scientific and technical reports became available, and providing access to this content challenged both the mechanized systems and the documentation methods of the time. In the Cranfield experiments (Cleverdon et al., 1966; Cleverdon, 1991) a comparative evaluation was performed regarding several existing indexing and classification methods. A laboratory setting, essentially, was introduced in these experiments in order to study the performance of index languages in isolation (Cleverdon, 1991). The experiments formed the prototype of future IR testing to come: test questions were collected from test persons; a document database was built; and intellectual relevance decisions were acquired so that the correct solution for each retrieval task was known beforehand.

Test collection-based evaluation experiments[9] in this thesis refer to the traditional Cranfield-style information retrieval evaluation performed in a controlled laboratory setting. The fundamental purpose of an IR system is to help its users find information contained in documents through string matching. This is challenging because natural language is rich and complex and it allows searchers and authors to express the same ideas in many different ways. Moreover, as vast amounts of text need to be searched, this sets demands for the data structures and algorithms used in order to allow rapid searching. Finally, individual users may have different opinions regarding what should be retrieved (Voorhees, 2007) and how to approach it.

The traditional methodology abstracts away details of particular tasks and users. Instead, an abstracted more general retrieval task can be solved in the test collection

[9] The expression traditional laboratory model of IR evaluation (Ingwersen and Järvelin, 2005, p. 2) refers to the same idea.

environment. The test collection consists of three components: a set of documents (the document database); a set of requests (topics); and a set of correct answers, i.e., the relevance judgments. Ideally, the relevance of every document in the database is known with respect to each pre-defined request. The research problems studied in such environment are related to various document and request representations and/or the methods of matching them. An important feature of test collections is that they are reusable. Given a test collection, researchers can quickly compare the effects of alternative IR approaches. The goals of research in test collection-based evaluation entail gaining theoretical understanding of the basic problems of IR (document and query representation, matching); developing theories and methods to deal with these problems; and developing matching methods successful in practice to retrieve relevant documents for users based on their requests (Ingwersen and Järvelin, 2005).

The scope of the experiments can be characterized in terms of types of experiments, test collections, requests, and performance measures used (Ingwersen and Järvelin, 2005). In a typical TCE experiment a set of topical queries is run as a batch-mode experiment. The test collection is a free-text collection, consisting of, e.g., newspaper articles; the requests are topical; and the relevance judgments are made by individual judges acting as relevance assessors. The performance measures typically take into account the quality of the retrieved result (a ranked list or set of documents) and effectiveness is measured in terms of available single-query metrics, not assuming several queries per single topic, and averaged over the set of topics or additionally expressed for each individual topic. The major performance measures[10] are based on recall and precision, e.g., mean average precision (MAP), and more recently, on measures like cumulated gain (CG) and its derivatives (discounted and session-based CG variants) (Järvelin and Kekäläinen, 2000; 2002; Järvelin et al., 2008).

[10] The effectiveness measures are discussed in Chapter 4.

## 2.2 Limitations of the traditional approach

The TCE model is justified based on the observation that character string-based matching of requests and topically relevant documents can be performed successfully. Natural language words extracted from the free text of the documents can be used as search keys to define topical requests because the existence of the words in the documents correlates with fair probability to the topical content of the documents they represent (Ingwersen and Järvelin, 2005).

The TCE model has been utilized extensively in IR, e.g., in various tracks of TREC (Voorhees and Harman, 2000; Voorhees, 2001, 2007). As a whole this methodology has been successful in allowing controlled studies regarding the performance of IR methods and systems based on their ability to find topically relevant documents. It has helped in laying the foundation to the technical solutions manifesting in the present day commercial IR systems, including many Web search engines (Voorhees, 2007).

Table 1 recaps the limitations of TCE discussed in the introduction, relates them to real life, and positions individual Studies I-IV of the thesis in relation to them.

*Table 1.* Typical assumptions in traditional TCE (column 1), in real life (column 2), and in Studies I-IV (columns 3-6).

| | **Traditional TCE** | **Real Life** | **Study I** | **Study II** | **Study III** | **Study IV** |
|---|---|---|---|---|---|---|
| **L1 User features** | No explicit user modeling | Various users, tasks, and situations occur | Explicitly modeled user behavior (relevance feedback) | Explicitly modeled user behavior (relevance feedback) | Negative higher-order relevance aspect | Explicitly modeled user behavior (direct query modifications) |
| **L2 Single versus multiple query** | Single; more than one in case of relevance feedback | Multiple queries if needed (real interaction) | Two: a simulated RF query modeled | Two: a simulated RF query modeled | Single | Multiple queries, direct query modifications |
| **L3 Nature of topics and queries** | Well-defined topics;<br>Often verbose queries | Well or ill-defined topics (user-given);<br>Often short queries | Well-defined topics;<br>Verbose queries (initial and RF) | Well-defined topics;<br>Verbose queries (initial and RF) | Well-defined topics;<br>Verbose queries | Well-defined topics;<br>Extremely short queries |
| **L4 Nature of relevance** | Static; topical, judged by an external assessor; often binary | Dynamic higher-order relevance affected by the situation, task, and the individual | Static; topical; graded relevance | Static; topical; graded relevance | Static; topical, limited higher-order relevance; graded and negative relevance | Static; topical; graded relevance |
| **L5 Document independence** | Yes | No; cumulating and redundant information matter | Yes | Yes | Yes | Not applicable |

| | | | | | | |
|---|---|---|---|---|---|---|
| **L6 Evaluation issue** | Various measures used; up to top-1000 documents retrieved evaluated | User-determined; often focus towards few top ranks | Focus towards high recall (top-1000 documents assumed); mean average precision | Focus towards top ranks; relevant information cumulates | Focus towards top ranks; relevant information cumulates; the importance of avoiding non-relevance | Exactly one (highly) relevant document required; top-10 documents inspected per query; binary success measure |

Table 1 makes it apparent that many assumptions of the traditional laboratory experiments (the first column) and observations from real life (the second column) are diametrical opposites. In the remainder of the present chapter we will discuss the six limitations of Table 1 individually.

### 2.2.1 Limitation 1: No explicit user modeling

In traditional TCE there is no *explicit user modeling*. Users with their tasks and situations are effectively abstracted away. This is both an advantage and a disadvantage (Voorhees, 2007). The experiments do not take into account an individual user having a particular cognitive state, experiencing learning effects during retrieval, possibly redefining the retrieval task, and who has a dynamic view regarding relevance during the search. On the other hand the strength of the model is to ignore variation regarding the users, tasks and situations. These attributes are not needed to study the limited goal of how well various representations and matching methods work in retrieving or ranking topically relevant documents (Kekäläinen and Järvelin, 2002a).

In real life, searchers have varying cognitive states (Belkin, 1980) while performing different kinds of work tasks. Consequently, different kinds of searching behaviors emerge (Stenmark, 2008) and affect the actual result the users will achieve. Interaction may involve user learning, problem redefinition, and changing relevance criteria (Bates, 1989; Kekäläinen and Järvelin, 2002a) which may be difficult to model. Yet it would be desirable – at least in principle - to consider different users and usages of systems explicitly in test collection-based IR experiments.

### 2.2.2 Limitation 2: Single query per topic

In traditional TCE batch-mode experiments a *single query per topic* is used and consequently the system's performance is expressed by averaging the results for such queries, over a set of test topics. This is justified because the systems should be rewarded from a good one-shot topical performance based on user's query as input - the systems cannot read the user's mind (Kekäläinen and Järvelin, 2002a). Yet, all in all, one query per topic situation is implicitly modeled (Kekäläinen and Järvelin, 2002a).

In real life when users operate IR systems, interaction[11] is the key element (Swanson, 1977; Belkin, 1980; Bates, 1989). It is common that the user issues an initial query and inspects some (top-N) documents retrieved. Spink (1997) observed[12] that in 40 % of the total interactive feedback occurrences a query was followed by relevance judgments before a modified or reformulated query or another command was entered. When an insufficient number of relevant documents is observed, the user may adapt via launching a modified query[13] or utilizing relevance feedback (if available). Such query modifications may be in fact unavoidable because even if the query does describe the topic well, it may have several interpretations (see, e.g., Sanderson, 2008) and retrieve documents not serving the particular need of the user. The process of launching queries and browsing their results is iterated until the searcher is satisfied or gives up. Such an iterative process is fundamentally different compared to the traditional view.

### 2.2.3 Limitation 3: Well-defined topics and queries

In traditional TCE *well-defined information needs* and *relatively verbose queries* are typical (Jansen et al., 2000). Well-defined topical needs allow making relevance

[11] *Interaction* can be understood in IR as sequences of events occurring in various connected levels. Surface level interaction encompasses a user dialogue with the system, including searching, matching, browsing, providing relevance judgments, feedback, and so on. On the cognitive level the user considers retrieved data as cognitive structures. On the situational level the user interacts with the task at hand which is producing the need for information. (Saracevic, 1996a)

[12] Real users performing mediated retrieval in a Boolean environment were observed.

[13] Users have been observed to resort to rapid, multiple query attempts (even if no topical redefinition takes place) even under heavy time pressure (Järvelin et al., 2008), instead of continuing browsing the initial retrieved result at extended lengths. Multiple attempts are typically carried out by making small query modifications by changing, adding or subtracting terms (Jansen et al., 2000).

judgments and constructing the recall base.  Using “reasonable queries” is understandable assuming the implicit presupposition of using a single query per topic in tests:  failing queries do not make a lot of sense from the point of view of comparing the effectiveness of IR techniques.

In real life, on the contrary, the specifiability of the information needs forms a continuum (Belkin, 1980). The information need of the searcher may not be well-defined; it may change (Ruthven and Lalmas, 2003), and even if it is well understood it may defy description as a query (Belkin, 1980).  Moreover, in real life a query lacking major facets of the underlying information need may be the preferred choice of the user. Such a query can be justified if it serves its purpose by leading to a *good enough* result, while simultaneously minimizing the effort to construct expressions[14] and inventing search words suitable as topic descriptors.[15]

In fact, it is realistic to assume that IR system users prefer using short queries - from one to three words, and often only one word (Ruthven, 2008) – instead of constructing verbose queries.  An analysis of thousands of queries posed by Internet search service users showed that the average number of terms used in a query was only 2.21 (Jansen et al., 2000) and even smaller, 1.45 terms, in a study by Stenmark (2008) focusing on intranet users.  Vakkari (2000) analyzed the relationship between students’ problem stages and search tactics in a longitudinal study, and observed that the number of search terms varied from 2 to 5 in the early stages of preparing a research proposal, and in the later stages from 3 to 11.[16]

We conclude by stating that as the queries used in IR tests are the basis of characterizing the effectiveness of IR techniques compared, it would be preferable to use realistic queries (and sequences of queries) in test-collection based experiments, as far as this is possible (see Kekäläinen and Järvelin, 2002a; Ingwersen and Järvelin, 2005).

[14] Ruthven (2008) notes that people may prefer short and *unstructured* queries.

[15] Using short queries may make sense even if long queries would provide a better retrieved result. Shorter queries may be more efficient in terms of communicating with the system. Azzopardi (2009) observed that the *change* in total performance divided by the change in query length was maximized when the query length was two terms - implying diminishing returns for the subsequent added terms.

[16] *Library and Information Science Abstracts* (LISA) database and a search system with Boolean operators were used in the study.

### 2.2.4 Limitation 4: Topical relevance

In traditional TCE topical relevance[17] forms the basis of evaluation. Also, traditionally, binary relevance judgments have been used, often with a low relevance threshold (Sormunen, 2000). Its advantages include simple performance calculations; low relevance assessment costs; and maximizing the number of relevant documents for attaining stable effectiveness measures (Sormunen, 2002). Although the general topical relevance criterion does not take into account the individual state of knowledge of a user or situational factors, retrieval methods can be compared based on the limited task of retrieving topically relevant documents (Kekäläinen and Järvelin, 2002a). The topical viewpoint is useful because it allows studying how well the system helps the user getting access to the subject material he needs, while also limited because in the ideal situation it would be desirable to measure the positive impact the IR system has for the particular user considering his search situation as a whole (Hersh, 1994).

In particular, one may question whether the lowest relevance threshold should be used in evaluation (i.e., marginally relevant documents are accepted as relevant) as the standard practice has assumed, especially in such collections in which the marginally relevant documents, by definition, do not convey any useful information (Sormunen, 2000).[18] The relevance grade reflects the amount of topical information “in the document” as judged by a human relevance assessor[19] (e.g., a non-relevant document or a marginally, regularly, or highly relevant document observed). Using graded judgments is justified because topically highly relevant documents can be recognized reliably (Sormunen, 2002; Vakkari and Sormunen, 2004).[20]

In real life beyond-topical relevance factors may be significant for the user, e.g., time pressure (related to situational relevance) and lack of accomplishment (related to motivational relevance) may affect the user’s behavior during interactive search

[17] Topical relevance can be defined as the relation between the topic (subject) expressed in a query and topic covered by information objects (Saracevic, 2006).

[18] Recently, Scholer and Turpin (2009) have suggested that marginally relevant documents should be grouped with non-relevant documents - not with relevant documents.

[19] See Sormunen (2000), p. 63, and Ahlgren (2004), pp. 164-165 for instructions for the assessors.

[20] Yet individual judges constructing graded recall bases may utilize different thresholds between relevance categories even if identical judging instructions are given to them. Moreover, end users may have different relevance profiles: not only may they have higher or lower criteria for relevance but also other user features like gender and age may affect the documents’ perceived usefulness (Scholer and Turpin, 2009).

process (Saracevic, 2006). Moreover, one may argue that even if the aboutness of the text is stable, the user's interpretations and thereby the perceived topical relevance may change during interaction (Cosijn and Ingwersen, 2000).

### 2.2.5 Limitation 5: Document independence

In traditional TCE the relevance of the individual documents must be judged independently from each other during the construction of the recall base of the test collection (see, e.g., instructions for the assessors in Ahlgren, 2004, pp. 164-165).

In real life, on the contrary, as Beaulieu (2000) states, individual documents are likely to be judged in relation to other retrieved documents. Moreover, the judgment process itself may lead to the reassessment of the initial relevance criteria themselves. A person evaluates documents in terms of his current state of knowledge, which may change upon receipt of information (Belkin, 1980). Also during the evaluation of the retrieved result the possible information overlap[21] in the documents retrieved is normally ignored (Kekäläinen and Järvelin, 2002a).

### 2.2.6 Limitation 6: Challenges of traditional evaluation

In traditional TCE incomplete evaluation measures like recall and precision are used which focus on the quality of the *result* retrieved. They exclude aspects related to the interactive search process. However, even if only the retrieved result is taken into account, it would be desirable from the user point of view to consider the part of the retrieved result the user actually sees. Therefore the documents in top ranks deserve special attention in interactive test collection-based evaluation. Users may differ regarding the lengths of documents sequences they are prepared to browse.[22]

[21] Discounted cumulated gain (Järvelin and Kekäläinen, 2002) models the phenomenon that the value of relevant documents seen later during retrieval is diminished partially due to redundancy (overlapping information) in *distinct* documents, thus considering the document dependence issue (Järvelin and Kekäläinen, 2002). It is currently an open question how to solve the problem of evaluating query sequences assuming that the user may purposefully select the *same* document multiple times using *varying* relevance criteria as search process evolves (Azzopardi, 2007).

[22] Top ranks are in the focus in traditional TCE evaluation in measurements like P@10 or nDCG@10. However, rank-wise browsing within a multiple query search process is not a part of the traditional TCE model because there is no concept of interactive search sessions and multiple queries.

The utility of a retrieval system could be defined not only in terms of how much the user gained in terms of useful information - but also costs or frustrations (Korfhage, 1997; Yang et al., 2007). Different users may have varying levels of satisfaction in receiving relevant documents and varying tolerance for frustration in receiving non-relevant documents (Korfhage, 1997), affecting their interactive behavior. It would be desirable to take such user features into account during evaluation. Such varying user behavior can be modeled during interaction simulations.

## 2.3 Bringing real life features into the model

From within the traditional model itself we cannot answer what are the consequences of abstracting away real life attributes (the limitations discussed above). Recent empirical tests involving real users have shown discrepancy between non-interactive batch evaluation results and interactive user evaluation results. Hersh et al. (2000) showed that the weighting scheme giving the maximum improvement over the baseline in non-interactive batch evaluation did not do so when real users performed a simulated task. Turpin and Scholer (2006) observed no significant relationship between the search engine effectiveness measured by mean average precision and real user success in a precision-oriented task. In a recall-oriented task a statistically significant but weak relationship was observed. Turpin and Hersh (2001) observed that a superior system to the baseline (in batch evaluation, measured by mean average precision) was not superior in an interactive situation.

Real users are able to successfully compensate for the performance differences by using interaction - e.g., by issuing more queries and reading more documents. Smith and Kantor (2008) observed that users of degraded systems were as successful as those using a non-degraded system, and suggest that they achieved this by altering their behavior. In the current thesis, our point of departure is the traditional single query per topic assumption which we expand via multiple query simulations. This is a relatively new way of approaching laboratory-based IR.

When the limits of the traditional TCE model are expanded through simulations, we need to model real life interactive user behavior and consider evaluation which is justifiable from the user point of view. The next chapter will explain our simulation approach. We will discuss real life IR interaction issues and bring them explicitly into the model in order to perform simulations. We will explicate our responses regarding the six limitations discussed earlier and suggest some solutions to them by using simulated experiments. The evaluation issue will be discussed separately in Chapter 4.

# 3. Simulating IR interaction

## 3.1 Modeling and simulation activity

A simulation can be defined as a symbolic model of a real-world situation created for the purpose of studying real-world problems (Adams and Rollings, 2007). As a starting point some *system* (physical or conceptual) is considered which consists of a collection of interacting entities producing some form of behavior. This behavior can be observed over an interval of time. A *model* is a representation or abstraction of the system in some form other than itself and acts as a surrogate for the system. *Simulation* consists of experimentation using the model (Birta and Arbez, 2007). Modeling and simulation are motivated by gaining insight into the features of system's behavior to provide deeper understanding, or to serve a more practical goal, e.g., to make changes in the system. Modeling activity is concerned with developing a specification for behavior generation usable as a vehicle for experimentation. An appropriate model needs to be considered in relation to the problems to be solved. To have validity, a simulation must represent relevant features of the real world (or any kind of simulation target) as closely as possible though aspects represented may be simplified or abstracted out (Adams and Rollings, 2007).

## 3.2 Extending the traditional approach

In Section 2.2 we presented some implicit limitations of the traditional TCE approach. Next we will discuss performing user simulations[23] in case of a multiple

[23] Ahlgren (2004) simulates the *user* dimension in a traditional single query setting by modeling patient and impatient users giving large or small differences regarding the values of the documents belonging to various relevance levels.

query situation[24] – a dimension so far lacking from the mainstream of the test collection-based IR studies. We will list our responses regarding the six limitations discussed in Section 2.2, and then discuss our simulations in more detail.

### 3.2.1 Response 1: Modeling user behavior

In the present thesis we will simulate explicitly the user dimension. We will model varying user (interaction) features and vary them systematically. Our goal is to simulate plausible interactive user behavior, justify it, and to demonstrate the effects of various "what if" scenarios regarding user behavior. User simulations performed in the laboratory allow studying systematically the effects of various interactive search strategies. If real users were used instead, it would be problematic to have control over specific types of user strategies[25], avoid learning effects, and support repeatability of experiments.

### 3.2.2 Response 2: Allowing multiple queries

One of the main factors in our thesis is that we focus on the *multiple query* aspect. Instead of implicitly assuming single query processes, we model scenarios where multiple queries are launched in a search session as a sequence. Using terminology by Järvelin et al. (2008) we call the topical multiple-query sequences *sessions*. We will focus on selected interactive user features in our simulations. In both RF and SS simulations two basic attributes are considered: the patience to browse the (initial) retrieved result (*B*) and the threshold to accept documents from various relevance levels as relevant (*R*). Additionally, in case of RF simulations we take into account the patience of the simulated searcher to collect feedback (*F*). In case of SS simulations we vary the specific prototypical query modification strategy (*SS*) used to reformulate queries in case a particular query failed. We will justify the selection of these attributes based on earlier studies on user behavior.

[24] White et al. (2005) simulate the *time* dimension (implicitly) by modeling a searcher observing sequences of document representations (e.g., a title or a top-ranking sentence) and study how well various implicit feedback models learned the term distribution across the relevant documents and helped to improve search effectiveness.

[25] For example, people may be reluctant to use RF (Dennis et al., 1998).

### 3.2.3 Response 3: Traditional topics; query types

The topics we use in our simulation experiments are traditional (static, well-defined topics), thus, our simulated user does not switch topic. The queries used in the tests are constructed using natural language keywords. We use both *long queries* (in RF simulations, Studies I and II) and *short queries* (in SS simulations, Study IV).[26]

### 3.2.4 Response 4: The role of higher-order relevance

Even accepting that static topics are used in simulations, truly user-based simulation would require that the searchers' relevance assessments are used instead of a recall base. Our study is limited by the fact that we will utilize traditional topical relevance conception in our simulation experiments, because the relevance data available in the test collections contain graded relevance judgments based on the amount of topical information in individual documents[27].

In real life higher-order relevance aspects (cognitive, situational and motivational) are present during the searching process. In our simulation experiments we assume that the patience of simulated users may be affected in various situations by factors like time pressure (situational level of interaction). We assume that the user's patience and impatience translates to surface level behavior so that the user's tolerance to browse the retrieved result and give relevance feedback may vary. In Study III we will also discuss the concept of negative higher-order relevance having desirable properties from the point of view of visualizing user experience. We will discuss the higher-order relevance aspect separately in Study III.

[26] Query formulation is explained in more detail in Sections 3.3.4 and 3.4.4.

[27] We do *not* attempt to reach in our simulations the level of cognitive user-centered approach and consider how well the user and the retrieval mechanism interact under real-life operational conditions (Borlund, 2000).

### 3.2.5 Response 5: Document dependence and independence

Because our simulations are based on a standard test collection, we do not intend to solve the *document independence* problem. In Study III we discuss how the discounted CG measure allows taking into account user-internal features of relevance. Partially due to acknowledging the document dependence problem[28] we will focus in Study IV on the special case of such a simulated user who wants to find only *one* (highly) relevant document, thus avoiding the document independence problem.

### 3.2.6 Response 6: Challenges of interactive evaluation

As discussed earlier, when the effectiveness of the retrieved result is evaluated in interactive IR from the user point of view, it would be desirable to consider the part of the result that the user actually sees. Therefore, knowing that the patience of the users to view the retrieved result varies, the number of *top ranks* considered during interactive session deserves attention.

Jansen et al. (2000) analyzed viewing behavior of Internet searchers and observed that users did not frequently browse the results “beyond the first page or so” (assuming ten documents per page). The mean number of pages examined per user was 2.35 and most users did not access any results past the first page. The authors argue that using a classical measurement of precision any search results beyond rank 10 would be meaningless for most users.

Our basic idea is to relate the user patience modeled to the length of the document sequence inspected during evaluation. We will discuss the document sequence formation problem (the ordering and length of the sequence), and how to evaluate the sequence, in case of RF and SS simulations in Chapter 4.

[28] E.g., the relevant information in relevant documents retrieved may be overlapping.

### 3.2.7 From systems to models

Our model specification begins with the description of plausible interactions within the system under investigation. IR systems can be considered as dynamic real life systems where interaction takes place in time. To justify any algorithmic simulation of (surface level) interactions, we will first discuss user interactions in the cognitive level. We will abstract a sequence of events in time as steps taken by the searcher – in particular, rank-wise browsing and feedback. We will characterize our IR interaction simulations as four phases:

1. *System characterization*
2. *Verbal model description*
3. *Formalized model description*
4. *Experimental procedure*

Next we will characterize the real life system assumed in the separate cases of RF and SS. Then we will verbally describe our RF and SS models, which are simplified abstractions of the system.[29] This is followed by a more formalized version of the model. Last, the experimental procedure is described.

## 3.3 Relevance feedback simulations

### 3.3.1 System characterization

An initial understanding of relevant features of the plausible real life RF interaction (i.e., *system*[30]) must be developed to form the basis for a valid RF model.

We assume a real life situation where a person is interacting with a query-based search system. The first query acts as an entry to the search system followed by possible subsequent phases of browsing and query reformulations (Marchionini, 1993). Relevance feedback is given based on the result retrieved: the searcher inspects some sequence of the documents retrieved (from top to bottom, see

[29] In course of the modeling process we present the research questions related to our point of view regarding the system.

[30] Even an "open" real life system characterization is itself a simplification (and a model). Obviously, it is possible to pay attention selectively to various aspects of the system in more detail in relation to specific research questions to be solved - thus leading to different models.

Joachims et al. (2005)) to identify which documents are relevant and which are not (e.g., Spink, 1997; Efthimiadis, 1996). Information in the relevant documents retrieved and seen by the user can be utilized for producing a *modified query* (a new search command input).[31] The modified query will (hopefully) be closer to what the user desires (Rocchio, 1971b).

The RF process described above obviously has many dimensions. The number of documents inspected and selected for RF may vary. The feedback terms may be selected and weighted in many ways - by the searcher or by the system. Moreover, the users may have various criteria for a successful search.[32] It is also known that the textual characteristics of documents belonging to various relevance levels vary: highly relevant documents discuss a larger number of topical aspects and they use a larger set of unique expressions (Sormunen et al., 2001).

Based on these observations we will address the following research questions in Study I:

*1. How effective is relevance feedback if we consider various thresholds of relevance in evaluation?*
*2. How is the quality and quantity of the RF related to retrieval effectiveness?*[33]
*3. Can pseudo-RF successfully compete with the simulated RF?*

Simulations allow us to explore the limits of the effectiveness of user feedback based on using higher or lower relevance thresholds during collecting the feedback documents and during evaluation, and assuming higher or lower effort in collecting the feedback. In Study II we will address the additional research questions:

*1. How should we evaluate the effectiveness of simulated user RF considering graded relevance assessments?*
*2. How successful are various RF strategies?*

[31] The user may select feedback terms *himself* based on the documents, or the terms may be extracted *automatically* by the system. The idea of automatic term extraction is justified because relevant documents may contain useful terminology (Ruthven and Lalmas, 2003) and because users are able to identify relevant documents, especially highly relevant documents (Vakkari and Sormunen, 2004).

[32] A user might desire to retrieve, e.g., few highly relevant documents, or lots of mixed-level documents (Kekäläinen and Järvelin, 2001; Voorhees, 2002).

[33] Note that even if the user requires, e.g., highly relevant documents, he might purposefully accept documents as feedback based on a lower relevance threshold.

Next we will describe how our simulation approach allows answering these research questions in a test collection-based setting.

### 3.3.2 Verbal model description

The *simulated* RF process (its surface-level manifestation) is described next. The retrieval system returns a ranked list of documents as a response to the initial query. The simulated user giving RF will browse at most (some number) *B* documents retrieved, from the first rank onwards, and recognize at most (some number) *F* as feedback documents, in case they satisfy a quality criteria defined by relevance threshold *R*.

Browsing limit, attribute *B* above, is justified because the user's willingness to study retrieved sets is limited, and it may vary individually in different situations (*futility point*) (Blair, 1984).[34]

Willingness to provide feedback, attribute *F*, is needed as a separate attribute because even if the user is willing to browse through a long list (high value of *B*) he may give up collecting feedback after finding the first (or first few) relevant documents.[35]

Last, relevance threshold attribute *R* is justified, because the users may want to focus on giving highly relevant feedback (Kekäläinen and Järvelin, 2002a; Voorhees, 2001). We are interested in studying via simulations whether this is a good idea.

In case of pseudo-RF all *B* top documents observed are used as feedback. In all RF scenarios of the present study the feedback terms are automatically extracted from the RF documents and the RF query is formed and launched automatically (Harman, 1988).

[34] Some users avoid browsing the retrieved results beyond the first few documents retrieved (Jansen et al., 2000) before attempting another query, thereby making small values of B reasonable. On the other hand some users, e.g., patent searchers, may require high recall (Kando, 2000) and be willing to scan through long lists of retrieved documents, thereby making high values of B and high final evaluation lengths reasonable.

[35] This dimension is essential because in real life people may be reluctant to provide feedback (Ruthven and Lalmas, 2003; Jordan et al, 2006).

### 3.3.3 Model formalization

The *user model* used in Studies I and II is a tuple $M = <R, B, F>$ where

- $R \in \{1, 2, 3\}$ is the requirement of relevance (at least marginally / fairly / highly relevant documents[36] were accepted as feedback documents)
- $B \in \{1, 5, 10, 30\}$ is the willingness to browse documents (at most $B$ top documents)
- $F \in \{1, 5, 10, 30\}$ is the willingness to provide feedback (at most $F$ feedback documents), $F \leq B$

### 3.3.4 Experimentation procedure

We used a TREC database and retrieval system *InQuery* (Broglio et al., 1994) and its *sum* operator[37] to envelope the query keys. In RF experiments we simulated a user who is willing to formulate a verbose initial query by extracting and lemmatizing all words in the *title* and *description* fields of the topic description of the test collection. In real life the user's ability to identify query concepts and articulate them is affected by, e.g., his subject and search knowledge (Vakkari, 2002) and his willingness to produce search keys is limited (Jansen et al., 2000; Stenmark, 2008). As keywords may be ambiguous (e.g., "left") and result in more than one lemma ("leave", "left") all the lemmas produced were included within a synonym (*syn*) set producing an initial query of the form:

*#sum(#syn(key1, key2,…), #syn(… key n), …)*

Top 50 documents are retrieved using the initial query. Specific value combinations defined by the user scenario $<R, B, F>$ were used together with the

[36] The test collection utilized in all our simulations was the reassessed TREC 7-8 collection (Sormunen, 2002) containing 41 topics from ad hoc tracks, 528155 documents, and graded relevance judgments. There were on the average 29 documents per topic belonging to relevance level 1 (marginally relevant documents); 20 at level 2 (fairly relevant documents); and 10 belonging to level 3 (highly relevant documents) per topic.

[37] See Ahlgren (2004) for a detailed description of the InQuery's operators and ranking principle.

recall base to recognize feedback documents in case of each user scenario and each initial topical search.

The 30 best feedback terms[38] were extracted[39], using a RATF formula (Pirkola et al., 2002) from the set of feedback documents.

The expansion keys were added in form of a sum structure (below: keys *e1, e2,* …) at the end of the initial query, to form the final *feedback query* for each scenario:

*#sum(#sum(#syn(key1, key2,…), #syn(… key n), …) #sum(e1, e2, …))*

As the last step, each feedback query was run in the test collection, and its effectiveness was evaluated.

Our simulations included specific value combinations (triplets) of *R, B,* and *F.* We assume that real users vary greatly; in Study I we experimented using value combinations *<R, B, F>* where $B \in \{1, 5, 10, 30\}$, $F \in \{1, 5, 10, 30\}$ ($F \leq B$), and $R \in \{1, 2, 3\}$ thus producing 30 simulated RF scenarios. The effectiveness of the initial query and each simulated RF scenario was analyzed using mean average precision (MAP) under three separate evaluation criteria: stringent, regular, and liberal relevance thresholds. In Study II, we experimented using a subset of value combinations, namely *<R, 1, 1>, <R, 5, 1>, <R, 5, 5>, <R, 10, 5>, <R, 10, 10>, <R, 30, 30>*, where $R \in \{1, 2, 3\}$, producing 6 x 3 = 18 simulated user scenarios. The effectiveness of the scenarios was analyzed using cumulated gain (CG) at the top ranks in order to focus on the user viewpoint in evaluation.

### 3.3.5 Pseudo RF

In pseudo-RF (Study I) the feedback terms were extracted without human intervention from a set of top documents retrieved and added to the initial query. The number of documents used for feedback can vary, and there are many ways to select the keys. We experimented with document set sizes of 1, 5, 10, and 30, and

[38] Also in a recent study adding 30 keys was observed to be a preferred choice in RATF-based expansion compared to adding 10 or 20 keys (Järvelin, 2009).
[39] Details of this process, including the RATF formula, are described in Study I.

used the same method (the RATF formula) for extracting the terms as in case of simulated user RF.[40]

## 3.4 Session strategy simulations

### 3.4.1 System characterization

When real searchers *directly* modify queries (without RF), the process is inherently complex. The user's initial state of knowledge leading to the query may be muddled, and it may change during the search process. Even if the need is well-defined, it may be difficult for the user to express it as a searchable query. (Belkin, 1980) A query may fail for many reasons, e.g., it may miss pertinent terms or contain ambiguous or too broad query terms and therefore retrieve documents not serving the particular need (Järvelin et al., 2008). From our simulation point of view we pay attention to the fact that in real life it is common that the users prefer using short queries (Vakkari, 2000; Jansen et al., 2000), they may revise their queries, there may be a need for multiple query iterations (Belkin, 1980; Ruthven, 2008) and the users may avoid excessive browsing (Azzopardi, 2007). Moreover, e.g., in Web searching searchers may view surprisingly little information without examining pages before attempting reformulation (Lorigo et al., 2006).

Our starting hypothesis is that the user behavior observed does satisfy user needs in many situations. Therefore, we are interested in seeing how effective short query sessions[41] are in a simulated situation. This motivates our main research question in Study IV:

*How effective are sequences of short queries combined with impatient browsing, compared to using one long query and patient browsing?*

[40] The PRF case can be characterized as a 3-tuple <$R$, $B$, $F$> where R = 0 and $F = B$: in PRF the $B$ top documents are always used as feedback (thus $F = B$) regardless of their level of relevance (thus $R = 0$). As mentioned, we used browsing lengths B such that $B \in \{1, 5, 10, 30\}$.

[41] In our simulations a session spans a sequence of one to five queries, all of which are related to one particular topic. This differs from the time-based session conception given in Jansen et al. (2000).

We will use two separate relevance thresholds (liberal and stringent) in evaluation defined for the simulated user.

### 3.4.2 Verbal model description

We state that the simulated user will launch an initial query; the retrieval system returns a ranked list of documents; and the user will *browse* some of the documents retrieved. During this browsing the user will either succeed in observing (enough) relevant documents or fails to do so. At some specific (rank) point regarding each particular query the user will stop browsing and instead launch *another query* (also followed by subsequent browsing) – which may be done repeatedly – or the user will give up the entire session.

### 3.4.3 Model formalization

The user model described above may be encapsulated as a 3-tuple $M = <R, B, SS>$ where, in our experiment,

- $R \in \{1, 3\}$ where $R$ is the requirement for target document relevance (*liberal or stringent*)
- $B \in \{10, 50\}$ where $B$ is the user's willingness to browse documents (at most $B$ top documents)
- $SS \in \{\text{S1, S2, S3, S4}\}$ where $SS$ denotes session strategies[42] describing how the prototypical query sequences are formed.

Next we will explain the experimentation procedure used in session strategy simulations.

[42] In Study IV four strategies S1-S4 were compared, i.e., $SS \in \{S1, S2, S3, S4\}$, see Section 3.4.4.

### 3.4.4 Experimentation procedure

Searchers modify their queries interactively because the initial query may not be satisfactory (Fidel, 1985). We will simulate idealized session strategies based on query modifications observed in the interactive query data by Lykke et al. (2009).[43] To construct query sessions, we first had two groups of seven test persons each to intellectually analyze the 41 topics used in the experiment. Regarding each topic a printed topic description and a task questionnaire were presented for test persons. They were asked to select and think up good search words from the topical descriptions, and form various query candidates: (i) the query they would try as the first query; (ii) the query they would try secondly assuming the first query did not succeed; (iii) queries of various lengths ranging from one to three or more words.

Using these word sets we constructed systematically[44] topical *query sequences* for simulations based on idealized session strategies. We constructed three types of idealized short-query session strategies (S1-S3) and one traditional strategy based on one verbose query (the baseline strategy S4). We restricted our attention to a situation where the user browses at most the top-10 documents (*B*=10) retrieved (S1-S3) for each query, and the *success criterion* was to find *one* relevant document. Two relevance thresholds (*R*=1 or *R*=3) were used[45]. We measured how many 10-document sequences needed to be inspected by the simulated searcher in order to reach success for each topic. We allowed a maximum of five queries per topic; if all five queries failed, the topical session failed. Two sets of query sequences were created for each SS type based on data collected from the two corresponding groups of test persons – students and staff members.[46]

*Sequences of one-word queries*

In session strategy S1 *one-word queries* are tried out one-by-one as a sequence.[47] If the user encounters 10 consecutive non-relevant documents, he will move on to the next query. New queries are attempted until the topical session ends successfully, or

[43] See Section 2.1 in Study IV.

[44] See Section 3.3 in Study IV.

[45] R = 1 denotes that marginally (or more) relevant documents are accepted as relevant; R = 3 denotes that only the highly relevant documents are accepted as relevant.

[46] We did not include staff members who had an extensive background regarding the specific test collection (e.g., IR researchers who had used the test collection in their own research experiments).

[47] A one-word query consists of a single search term which is an unbroken string of characters (letters or digits with no space between).

the simulated user runs out of query candidates (and the topical session fails). In other words, the following topical query sequence is attempted in S1:

*key 1-> key 2 -> key 3 -> key 4 -> key 5*

Strategy S1 simulates a user who tries to minimize his effort regarding each individual query. This strategy is justified because one-word queries are common in real life and people may try them as sequences. We purposefully experimented with this "most extreme" strategy even though it seemed obvious that it may perform poorly.

*Incremental query extension*

In session strategy S2 the simulated user starts the topical session using a one-word query. In case of failure the user *adds* one word to the query[48] every time until the session succeeds or the simulated user runs out of queries. At most five queries[49] are formed also in this strategy.

*key 1 -> key 1 key 2-> key 1 key 2 key 3 -> key 1 key 2 key 3 key 4 -> …*

Strategy S2 simulates a lazy searcher who tries to cope with minimal effort and adds words one at a time in case of failure. This strategy was observed in the query session data (in 13 out of 60 sessions) analyzed in Study IV.

*Variations on a theme of two words*

In session strategy S3 the simulated user always uses three-word queries but the two first search keys are fixed and the *third key is varied*, in other words

*key 1 key 2 key 3 -> key 1 key 2 key 4 ->key 1 key 2 key 5 -> …*

[48] Unstructured *#sum* queries of *Lemur* retrieval system were used in all SS experiments because facet structure was not considered.

[49] We argue that 5 queries is a reasonable number. A simulated study involving real users performing simulated tasks observed that the length of the search sessions ranged from 1 to 11 queries with a mean of 2.85 and a median of 2 queries per session (Price et al., 2007).

This strategy was common in the query session data analyzed in Study IV (38 of 60 sessions). Also Vakkari (2000), who analyzed the relationship between the users' problem stages and search tactics, observed a pattern where the user retains one or two terms and varies one term by substituting it with a new term.

All three prototypical query strategies S1-S3 are justified because they were observed in empirical data by Lykke et al. (2009). Moreover, they are interesting per se because searchers may use such queries as building blocks during trial-and-error type search processes in real life. Testing them is well motivated in a multiple query situation even though in a single query situation they may not seem to be particularly interesting.

### *The baseline query*

Last, the baseline query strategy S4 consisted of *one* verbose query, in which case at most the top-50 documents were browsed by the simulated user, in 10 document chunks.

# 4. Evaluating simulated IR interaction

In this chapter we first explicate some problems related to evaluating simulated interactive information retrieval in test collections. We will describe some solutions to the problems and explain our RF and SS evaluations performed in Studies I – IV. We conclude this chapter by discussing statistical testing.

## 4.1 Evaluating interactive IR in a test collection

Interactive information retrieval covers a wide range of research related to studying diverse end users of information access systems. It is shaped by research on information seeking and search behavior, and by research on the development of new methods of interacting with electronic resources. (Ruthven, 2008) Currently, there is a lack of research activities in modeling interactive IR systems (Ruthven, 2008). Also evaluating them is currently a major challenge within IR (Azzopardi, 2007). The IR research community has recently acknowledged the need to develop simulated evaluation of interactive search scenarios (see, e.g., Fuhr et al., 2009; Azzopardi et al., 2010).

The concept interaction has been characterized in various ways in different fields like human-computer interaction and information seeking/searching behavior (see, e.g., Beaulieu, 2000). A necessary condition to interaction is that some form of *feedback* takes place (Spink, 1997).[50] When interactive IR is evaluated in a simulated setting based relevant documents[51] retrieved, two issues must be settled.

[50] In traditional IR model feedback refers to an automatic function of an IR system where the user's query is automatically reformulated by the system (Spink, 1997). In interactive models feedback is seen in a context involving both the system and the user. In the current study we assume that interaction manifests, as Spink (1997) describes, as a search activity which consists of cycles of interactive feedback loops incorporating the user and the system inputs and outputs, together with user interpretations and judgments.

[51] We use relevance judgments of the recall base of the test collection in simulations, not judgments of test persons who make their own relevance assessments. We also acknowledge, but do not attempt

First, how should we construct the sequence of documents which acts as the basis for evaluation (i.e., the *ordering* and *length* of the sequence)?[52] Secondly, once some document sequence exists, what kind of *measurement* should be used to evaluate interactive search process, taking some specific type of user into account? Next we discuss these two issues – the sequence construction problem (its ordering and length), and the measurement problem.

### 4.1.1 Constructing the document sequence

A starting point for IR evaluation is that the effectiveness of any retrieved document set or sequence can be evaluated as such (called *the total performance evaluation*) using appropriate effectiveness measures. However, it is problematic to evaluate documents retrieved as a result of some interactions this way, because the relevant documents already seen[53] by the user (or simulated user) may be moved upwards in the ranked list. One may argue that such re-ranking makes the evaluation appearing artificially better than it really is (called *the ranking effect*).

Various solutions to this problem have been suggested in the literature. In case of relevance feedback three classical solutions to the total performance evaluation problem include *traditional freezing*, *full freezing*,[54] or splitting the collection into *test and control groups*. (Chang et al., 1971; Salton, 1989; Ruthven and Lalmas, 2003)

In *traditional freezing* the ordering of the document sequence for the RF query is constructed so that the relevant documents retrieved earlier are frozen in their original ranks and the non-relevant documents retrieved earlier are removed from the collection. The ranks of the non-relevant documents are occupied by the items newly retrieved in a subsequent search iteration (the RF query) (Salton, 1989). This approach can be criticized as being somewhat artificial from the user point of view.

to deal with the fact that real user's knowledge state may change on receipt of information affecting the optimal order of presentation of texts for the user (Belkin, 1980).

[52] Regardless of the type of interaction leading the user to the document (e.g., relevance feedback, direct query reformulation, or following links) the user will encounter some specific document sequence. We may assume that only one document can be examined by a searcher at any one point in time. It is this sequence that determines the effectiveness as experienced by the user. (Azzopardi, 2009b) Therefore, evaluating interactive information access systems could be based on these sequences accessed through the course of some particular interaction (Azzopardi, 2007).

[53] Documents may have been retrieved for the topic by some earlier query.

[54] Table 1 in Study II illustrates various freezing approaches.

In *full freezing* all documents (also non-relevant ones) seen by the user during the initial browsing are frozen at their ranks. All the yet unseen documents returned by RF are placed into the following rank positions. This naturalistic approach to the ordering problem can be justified by stating that it imitates the viewpoint of a user who has in fact wasted effort in browsing some particular sequence of documents regardless of their relevance level.

In *test* and *control* group solution the document collection is first split into two sub-collections. The modified (relevance feedback) query is constructed using the feedback from the test group documents. In this situation the document ordering in the *control* group becomes unproblematic and the ranked positions of the documents retrieved using both the initial query and the modified query can be accepted as such (i.e., the total performance measure can be used). (Ruthven and Lalmas, 2003)

Regarding the *length* of the document sequence evaluated, the main principle in Section 4.2 is that the perceived utility is “the only utility that does the patron any good”. Therefore, one should consider in evaluation the quality of the retrieved result regarding exactly those ranks that the user will observe. (Cooper, 1973)

As users are different regarding their patience to browse documents during interaction, there cannot be a natural single answer to the question of how many ranks should be considered in evaluation (e.g., “the first 1000 ranks).[55] Therefore, if different users are modeled, one needs to justify the last rank position inspected in each case. For example, if we simulate a user who is willing to study only short lists of retrieved documents (per query), due to, e.g., time pressure (Price et al., 2007) we may focus on few top ranks only during evaluation. Note that mean average precision may include the implicit assumption that top-1000 documents will be inspected. Cumulated gain (CG), instead, sums the gain values from rank 1 to rank i (when i ranges from 1 to the last rank position inspected) [56] allowing the cumulated value inspected to varying last rank positions needed. We argue this type of evaluation approach is justified in user simulations where the rank-wise inspection of the retrieved result by the user is modeled – compared to the system-based view

[55] As an example Pollock (1968) describes an “officer” who is willing to inspect only the first three documents to find one relevant document, or differently, an “analyst” who might be content with a long list containing 400 documents. A desirable measure of effectiveness should take into account such different uses of the list when the length of the document sequence is considered in evaluation.

[56] E.g., retrieved documents turned to a list of gained values G = <3,2,3,-1,-1,-1…> will cumulate as a CG vector: CG = <3,5,8,7,6,5,…>. For formal definitions see Järvelin and Kekäläinen (2002).

manifested in measurements such as MAP. We will discuss this issue in the next section.

### 4.1.2 Measuring the sequence

Real users may have various goals regarding both the retrieved result needed (Scholer, 2009) and the properties of the interactive process itself (e.g., as un-explicated restrictions) (Järvelin et al., 2008). The primary goal of finding the desired objects during interaction may be affected by the need to find them in some particular way. For example, the user may prefer to avoid wasting effort both in entering query words, and in browsing the result list – and not only in the latter. The desire to optimize balancing between such efforts can be modeled by using simulations. Therefore, when the effectiveness of the simulated interaction is evaluated, the evaluation measures need to be justified from the point of view of the user modeled.

Different measures have be used in the traditional TCE to evaluate the retrieved *result* (typically the sequence of documents retrieved), including test-collection based relevance feedback studies including mean average precision (MAP)[57] and cumulated gain (CG) based measures. Once the *ordering* and the *length* of the document sequence have been decided, various measures can be used to measure the "goodness" of the document sequence based on the existence of relevant documents in the sequence.[58] Different measures will emphasize different properties of the sequence. Therefore, the selection of the measure depends on what user preferences the evaluator wants to study (Kekäläinen, 2005).

Evaluation methods will affect what one observes (Kekäläinen and Järvelin, 2002b). The relative effectiveness of retrieval systems may change when the basis

[57] In non-interactive retrieval situations well-known effectiveness measures for evaluating the quality of the set of documents retrieved include precision (P), which is the share of relevant documents among all documents retrieved (Rocchio, 1971a): P at specific rank (e.g., P@10, i.e., P based on the first 10 documents retrieved); average precision (AP) which is the average of the P values calculated at the rank of each relevant document (non-retrieved documents get the precision value zero); and the mean average precision (MAP) – the mean of AP values over multiple topics; or measures like cumulated gain or its rank-wise discounted versions.

[58] Traditional recall bases cannot address the fact that users may learn during the retrieval process. In real life even the same document can attract different judgments from the same judge depending on where it appears in the stream of documents observed (Azzopardi, 2009b).

of evaluation is changed (e.g., highly relevant versus at least marginally relevant documents are accepted as relevant) (Voorhees, 2001).

In the CG measure the *degree* of relevance of each document is taken into account as a gained value for its ranked position in the result list. The length of the document sequence evaluated may vary while the gain is summed progressively from the first to the last rank inspected (Kekäläinen, 2005). The principle of CG-based measures is that at the last rank it gives a single composite estimate of the search utility for the user after the search has ended.[59] The resulting single combined measure is directly user-oriented and it allows defining various weighting schemes related to different relevance levels depending on the user modeled, and it is intuitive showing at *each* rank the gain a user gets at a given cut-off point of the result set. The length of the document sequence evaluated can be related to the behavior of the simulated user (assuming more or less patient users).

Next we will explain how the document sequence construction and measurements were performed in our individual studies.

## 4.2 Evaluations performed in the individual studies

*Study I*

In this paper we measured the effectiveness of different simulated users in simulated RF when the quality and quantity of RF is varied, and when the user requires documents belonging to different relevance levels during the final evaluation.

In our first study we used a system-oriented measure MAP in evaluation. MAP does not take into account that degree of relevance may vary in documents (Kekäläinen, 2005) yet separate relevance thresholds can be used. We needed to consider how to measure effectiveness in interactive IR using graded relevance judgments. We measured MAP values for the RF and PRF scenarios using three distinct relevance thresholds.

The *ordering* within the document sequences retrieved for both the initial query and the RF query were kept as such (total performance evaluation). Freezing was

[59] Järvelin et al. (2008) introduces *session-based* CG, an extension to CG, which penalizes further queries in a multiple-query session because each new query formulation requires further effort.

not utilized in the first study because we wanted to compare the effectiveness of simulated graded RF to pseudo-RF as such. A problem with such an approach is that it allows reordering of the relevant documents observed. However, we may assume a user who is simply interested in the quality of the *final* search result – in such a situation reordering is not problematic. We wanted to directly compare the result of simulated user-given RF (when its quantity and quality changes) to the effectiveness of pseudo-RF, because there was no prior information about such a novel setting.

The *length* of document sequences used for evaluating the interactive sequences was top 1000 documents. This is the traditional approach which implicitly models users who are willing to dig deep down the list of retrieved documents. We used the traditional measure in our first interaction simulation experiment because our focus was on developing an interactive user model for RF simulations and MAP served as an established way to consider performance. It serves as a reasonable starting point in our formative research process considering interactive user simulations in test collections.

*Study II*

Here we continue to study simulated RF and use a subset of 6 simulated RF scenarios from the first study. Because "total performance evaluation" (i.e., no freezing) was used in the first study, we wanted to focus in this study on how the evaluation of effectiveness of simulated RF scenarios is affected (if it is), employing different evaluation assumptions.

The *ordering* of the document sequence in evaluation in Study II was based on *freezing all* documents – both the relevant and the non-relevant documents – seen by the simulated user at their ranks.[60] This is justified because the user has already wasted effort in browsing and inspecting the documents seen despite their level of relevance. Because the past sequence of retrieved documents will be left as it is, it is obviously an advantage not to browse sequences of non-relevant documents any longer than necessary – like in real life.

The *length* of the sequence is considerably shorter in Study II compared to Study I. We varied the patience of the simulated user both during the simulated feedback phase and during the evaluation phase. We measured the effectiveness up to the last

[60] See *freeze all* case in Table 1 in Study II.

rank position of interest (rank 10, 20, or 100, depending on the scenario). The sequence length used in evaluation in this study was related to users having varying levels of patience. In the impatient user scenario only the documents initially retrieved within the top 5 ranks could be used as feedback, and the user would browse altogether only the first 10 ranks. In the most patient user scenario we assume that relevant documents initially retrieved within the topmost 30 ranks are first used as feedback, and the simulated user will browse altogether the first 100 documents retrieved during topical search session.

The *measurement* of the sequence was based on cumulated gain (Järvelin and Kekäläinen, 2000). We used CG in evaluation because it directly expresses the values the user actually gets when he browses a document sequence of specified length while the documents may belong to various relevance levels. We assumed a steep weighting scheme[61] 0-1-10-100 because we modeled users who appreciate highly relevant documents.[62] We did *not* use discounting[63] because our scenarios explicitly determined the patience of the users to examine a specific number of top ranks retrieved. Therefore, as we knew that the simulated user *will inspect* the varying lengths of top ranks, we wanted to determine the direct cumulated gain he gets by inspecting these ranks.[64]

*Study III*

In this study we do not perform interactive simulations. We suggest taking into account both the user's gains *and costs*, belonging partially to the higher-order relevance[65] domain, and define an extension to the CG-based measures using

[61] In weighting scheme 0-1-10-100 non-relevant documents are given weight 0 by the simulated user; marginally relevant documents weight 1; fairly relevant documents 10; and the highly relevant documents weight 100. This is justified because the marginally relevant documents do not contain extraneous topical information by definition (Sormunen, 2002) whereas highly relevant documents contain lots of topical information and are valuable for the simulated user.

[62] These weights are used also in Kekäläinen (2005) and Järvelin et al. (2008). They model a user who values highly relevant documents one hundred times more than marginally relevant documents, and ten times more than fairly relevant ones. The weights are not based on empirical evidence, thus they represent a "what if" scenario regarding the assumed values.

[63] Discounting models the phenomenon that it is less likely that the user will ever examine the document due to time, effort, and cumulated information from the documents already seen. See Järvelin and Kekäläinen (2000).

[64] Discounting would have squeezed all the performance curves downwards and closer to each other.

[65] The traditional test collections are based on topical relevance, and the relevance assessments of the documents reflect the "amount of" relevant information "in the document". This mindset seems to imply that non-relevant documents should be given zero weights. However, in case of higher-order relevance, the relevance is a *relationship* entailing not only non-negative aspects.

*explicit negative* gain values.[66] They allow us to visualize the frustration inherent in encountering non-relevant documents. Negative higher-order relevance refers to the negative aspects of beyond-topical relevance, e.g., dissatisfaction, frustration, and uncertainty experienced by the user. We suggest that research should recognize the existence of negative user sentiments caused by observing sequences of non-relevant documents during the search process.[67]

Recognizing negative sentiments during browsing is inherently related to the idea of interaction. The avoidance of wasting effort in non-relevant documents encountered may cause the user to prefer attempting another query (thereby encouraging interaction) instead of continuing browsing. Positive-only values miss this phenomenon as they seem to suggest that the user might continue browsing at extended lengths. As it may be important for the users to *avoid* non-relevant documents, explicit negative values are useful as they allow making this aspect visible.

*Study IV*

In this study we simulated a user attempting to rapidly find one (highly) relevant document by performing direct query reformulations. The short sequences of documents inspected (10 documents) simulate an impatient (albeit quite realistic) user who wants to avoid non-relevant documents and reformulates the query immediately in case of failure instead of continuing browsing, while taking chances with using purposefully “obviously incomplete” queries. Therefore the *success measure* used was a simple *binary* success measure. Success was determined as being able to find *one* relevant document using distinct two relevance thresholds (either a marginally or a highly relevant document was required).

The *ordering* of the short sequences of documents retrieved (top-10 documents) by each individual query (and inspected by the simulated user) was left untouched.

[66] We suggest in Study III that negative higher-order relevance can be operationalized simply by allowing negative gain values in CG-based measures. Using such values will lead to CG and DCG curves which may alternate up and down, instead of only being able to go upwards, thereby allowing the researcher to recognize concepts like progression towards *success* (or failure); raising (and descending) expectations, turning points in sentiments, and (more or less severe) *frustration experienced* by the user while encountering non-relevant documents.

[67] Using positive-only values for CG (or DCG) will have the consequence of having the performance curves keep going upwards (and never descending – not even at very high ranks) which can be considered counter-intuitive from the user point of view.

In other words we accepted the ordering of the top-10 documents retrieved by each query and retrieval engine as such. A maximum of 5 queries were launched for each particular topic (and, correspondingly, at most five “pages” each containing 10 documents, were inspected for each topic).

The *length* of the document sequences inspected in this study was short because of the type of the searcher simulated. As explained earlier, in all session strategies S1-S4 (see Section 3.4.4) we were interested in seeing how many “chunks of 10 documents” the user must browse to succeed[68] - or whether he fails altogether for the topic. In case of short-query strategies (S1-S3) each individual query either succeeded or failed within the top-10 documents it retrieved. In case of the long baseline query (strategy S4) we inspected a maximum the top-50 documents, while observing how many chunks of 10 documents must be browsed to find the relevant document.

## 4.3 Statistical testing

The purpose of significance tests in IR evaluation is to find out whether differences observed in the effectiveness of the methods compared could have occurred by chance (Hull, 1993). In IR the setting is typically arranged so that the associations between the dependent variables, e.g., effectiveness measures of mean average precision, cumulated gain, or the ordinal of the successful query within topic, and the independent variables – in our case a specific type of simulated interaction - can be identified.[69] When the measured values for the approaches are compared, statistical tests are used to make a conclusion whether the differences between the methods are statistically significant.[70]

Non-parametric tests have less power to correctly reject null hypothesis than parametric tests, but their requirements on sample sizes and the distributions are not

[68] In the best case the relevant document would be located within the first 10 documents, and in the worst case no relevant documents are observed after browsing through 50 documents.

[69] In our studies the null hypothesis states that that there is no difference between the retrieval methods compared based on the effectiveness.

[70] Statistical testing acknowledges that the conclusion made is false with the specific probability. We are interested in avoiding type I error, that is, the wrong claim that “A is better than B”, although in fact it is not. It is also possible to err in the opposite direction and fail to acknowledge A being better than B although it is (type II error).

as strict as those of their parametric counterparts.[71] Therefore, non-parametric significance tests are often seen as a justified choice in IR experiments.

We used the Friedman test which is a non-parametric ANOVA version (Hull, 1993). We used the test in RF and SS simulations in Studies II and IV. The formative Studies I and III focused on constructing the view regarding the usage of test collections with interaction, and therefore statistical testing was not performed in them. We utilized Friedman test in the current study due to the properties of the data and because we compared more than two methods in our studies. The Friedman test calculates first whether significant differences overall between the methods are found. If such differences are found, a pair-wise comparison between different methods is done to show which methods differ significantly from each other.

The Friedman test seeks to verify whether *k* related samples come from the same populations, or populations with the same median (i.e., that the systems or methods compared are equivalent). The data are arranged in *b* rows and *k* columns, where rows represent units (e.g., *b* individual topics) and the columns represent the respective treatments (e.g., *k* matching methods compared). The scores of each row are ranked from 1 to *k*, and the test determines the probability that the rank totals for each treatment (matching method associated with the column) differ significantly from the values that would be expected by chance (Conover, 1980). The test compares the absolute values of the difference between the column-wise rank sums $R_i$ and $R_j$, where $R_i$ ($R_j$) is the rank sum for the *i*th (*j*th) treatment condition, $i \neq j$, to a critical *z* value (Siegel and Castellan, 1988, pp. 174-183).

In Study II the statistical testing was based on cumulated gain values for each topic. Both the final cumulated gain value and the averaged cumulated gain value from the first rank to the final rank were used. In Study IV the effectiveness was measured as the ordinal of the successful query for each topic within a topical session. Significant differences were observed between the interactive search approaches compared (RF strategies in Study II; SS strategies in Study IV) in both studies.

[71] Problems of statistical testing in IR experiments have been related to small sample sizes, the nature of test requests (selections instead of random samples), and to the distribution of recall values over queries (for discussion see Kekäläinen, 1999).

# 5. Summary of the individual studies

In this chapter we will summarize the four individual Studies and focus on the research results.[72]

## 5.1 Study I

This paper opens our test collection-based user simulation approach. We will approach user relevance feedback issue by simulations because experiments with real users are time-consuming and also problematic due to learning effects, repeatability and control. In this paper we will define a user model and use it to quantify some major interaction decisions involved in simulated user RF.

*Research questions*

In this study our research questions are:

1. How effective is relevance feedback if we consider various thresholds of relevance in evaluation?
2. How is the quality and quantity of the RF related to retrieval effectiveness?
3. Can pseudo-RF successfully compete with the simulated RF?

*Methods*

We performed a simulation by constructing and utilizing the user model <R, B, F>[73] which explicates some basic interaction decisions of the user. When document level feedback is used, the quality and quantity of this feedback may vary.

At this point we had no idea whether the user should use higher or a lower relevance level threshold during RF. We were interested in seeing whether it makes sense to

[72] We will refer to the result tables and figures in the original Studies and do not duplicate them here.
[73] See Section 3.3.3.

use “mixed quality” RF (accepting documents from all relevance levels) or demand “high quality” RF (accept only highly relevant documents) assuming that during final evaluation the relevance threshold can vary. Our simulation model (Section 3.3.3) allows using these attributes - varying the relevance threshold to accept documents as feedback (e.g., only highly relevant documents are accepted); the maximum browsing length regarding the retrieved documents studied by the simulated user (e.g., the first 10 documents initially retrieved); and maximum number of feedback documents given (e.g., using only the first relevant document observed as a feedback document).

Evaluation of the final retrieved set was measured after the feedback based on altogether 30 user scenarios (three different relevance thresholds during feedback multiplied by ten different value combinations of browsing and feedback efforts) using a classical precision-based measure MAP.

*Results*

For the first and the second research questions, our results indicate that RF can be effective at all three evaluation levels.[74] The best RF scenarios also clearly outperformed the PRF scenarios. Regarding the third research question, also pseudo-RF improved the initial retrieval.[75] Interestingly, when stringent relevance threshold was used in evaluation the best simulated user RF scenario clearly outperformed PRF, but instead when a liberal evaluation threshold was used, the performance of the user scenarios in RF was close to the PRF results.

This simulation study left us with the open question of how one should perform evaluation during RF considering the user viewpoint. For example, if we accept MAP evaluation (based on the top-1000 documents retrieved) and allow reordering of the relevant documents observed, and assume a user demanding highly relevant documents during final evaluation – as we did in Study I – then using high quality RF is superior compared to using mixed quality RF.[76] However, what happens during the RF scenarios to the rank-wise results, in particular, when reordering of the seen result is not allowed? We will continue studying these issues in Study II.

[74] See Tables 5-7 in Study I.

[75] See Table 8 in Study I.

[76] See Study I, Table 5.

## 5.2 Study II

In this study we focus on user-oriented evaluation issue using cumulated gain-based evaluation (Järvelin and Kekäläinen 2000, Järvelin and Kekäläinen 2002). This approach allows *rank-wise* inspection of the results and it is directly user-oriented. As Azzopardi (2007) argues, IR evaluation could be based on those and only those documents (within the ranks of any sequence inspected) that the assumed user will actually see. Therefore, cumulated gain-based measures are especially well-suited for simulations. Different simulated scenarios explicitly model users who are more or less patient to browse the lists of retrieved documents. The gain values can be presented regarding the appropriate ranks – e.g., only the first top ranks in case of impatient users, and longer ranked sequences in case of more patient users.

*Research questions*

In this study two research questions were addressed:

1. How should we evaluate the effectiveness of simulated user RF considering graded relevance assessments?[77]
2. How successful are various RF strategies?

*Methods*

We used full freezing of the results in which all documents browsed (i.e., seen by various kinds of simulated users) are frozen at their ranks. We modeled *impatient* users allowing a small browsing window size during the relevance feedback phase (at most the top-5 documents). In case of such impatient users one may argue that it makes sense to consider during final evaluation only those documents that the user is assumed to see (Azzopardi, 2007). Therefore we assumed in evaluation that the simulated user will inspect only the first 10 documents retrieved.[78]

We also modeled a *moderately patient* user assuming he may examine more documents, and give more feedback (at most top-10 documents during the feedback

[77] Binary relevance has been used historically in test collection-based RF experiments. In such collections the idea that a user might demand documents from a specific level of relevance (e.g., highly relevant documents) but purposefully give, e.g., lower level feedback, is not obvious.

[78] See Figures 1-3 in Study II.

phase); therefore in evaluation we assumed that the simulated user will inspect the first 20 documents retrieved altogether.[79]

In the *patient* user case we model a user who may first examine at most top-30 documents during feedback phase. In the evaluation phase we assumed altogether up to the top-100 documents inspected by the user.[80] In all cases we varied the relevance threshold used to accept documents as feedback.

*Results*

The overall result can be summarized as follows. Despite full freezing the RF scenarios generally performed significantly better than the baseline scenario measured by final CG. However, significant improvements were not found in one case: when the user set unrealistic demands for the combined quality and quantity of the RF documents. When the demand for feedback quality is increased, there is less such feedback available.

Compared to Study I results, the general perception regarding the simulated scenarios changes. The highly relevant documents seemed to be very effective when reordering was allowed and evaluation was based on MAP (Study I). However, if all the documents seen are frozen (i.e., reordering of relevant documents is not allowed), and *rank-wise* inspection of the result (CG) is assumed, it makes sense to use mixed-level feedback (Study II), especially for impatient and moderately patient users. If the user is very “picky” and only accepts highly relevant documents as feedback, then no such feedback may be available in small browsing windows. In such cases no improvement can be made regarding the baseline.

For very patient users it makes sense to give lots of feedback - of *mixed* quality (i.e., using a low relevance threshold) - although compared to the high quality feedback scenario the final gain values at the last rank are close to each other. The general conclusion regarding mixed-quality RF is supported by topic-by-topic results: if a small browsing window is used in collecting feedback, high quality feedback may simply not be available.

[79] See Figures 4-6.

[80] See Figures 7-9.

## 5.3 Study III

In this study we continue to focus on the problem of user point of view during evaluation. We utilize a dichotomy *topical* versus *higher-order relevance* where the latter refers to the beyond-topical relevance criteria (Kekäläinen and Järvelin, 2002a). Cumulated gain (CG) (Järvelin and Kekäläinen, 2000) allows rank-by-rank inspection of the retrieved result. Its discounted version (DCG) (Järvelin and Kekäläinen, 2002) additionally allows modeling aspects of higher-order relevance by diminishing the value of relevant documents at later ranks.[81]

The novel concept of *explicit negative higher-order relevance* introduced in Study III underlines the importance of avoiding browsing non-relevant document sequences from the user point of view. High negative values for non-relevant documents can be utilized to model a user who does not tolerate well non-relevant documents (because he, e.g., gets tired easily, or is impatient or busy due to his work task or search strategy). The view developed in this study led us to use very short browsing windows repeatedly in Study IV.

*Research questions*

In this study our research questions are:

1. What is negative higher-order relevance and what is its justification?
2. How can we operationalize negative higher-order relevance?
3. What are the consequences of allowing explicit NHOR in IR evaluation?

*Methods*

We focus on non-relevant documents in the observed document sequence. Studies on real users show that users often seem to have “short attention span” in formulating queries (short queries are popular), reformulating them (small modifications are common), and toward browsing the retrieved results (only the top-10 documents or the first results page is observed). The users also often require good documents but not necessarily many of them – even one highly relevant document may suffice. One way to consider the worth of non-relevant documents -

[81] This is justified because the information already cumulated, redundancy, and effort render relevant documents at later ranks less valuable for the user.

the conventional approach – is to assign zero values to documents containing a zero amount of (useful) topical information. In other words, the user modeled is indifferent toward them. When the results of interactive user studies are considered, this seems to be far from the truth. As a solution we explain and justify using direct negative gain values in cumulated gain-based evaluation to operationalize explicit negative higher-order relevance.

*Results*

The novel concept of negative higher-order relevance introduced in Study III is useful as it makes visible the fact that not only users want to find relevant documents (to maximize gain) but they specifically may want to avoid non-relevant documents, e.g., as part of minimizing costs. Therefore the user, in order to stop further progression in an obviously "wrong direction", will rather discard browsing and try something else instead (e.g., a query re-formulation).

Our main result is that the concept of bi-directional gain/cost invites one to recognize novel concepts which traditional measures do not suggest.[82] The most imminent is *bi-directional progression* – towards success or failure. Bi-directional progression also raises questions about *maximum utility* and the questions related to the concepts of *failure* and *success*.[83] In real life the users normally stop browsing after observing sequence of non-relevant documents (although in some situations users may be patient and browse very long lists of retrieved documents). The considerations presented in the third paper, together with the earlier two papers, lead to the setting of the fourth paper, which will be explained next.

## 5.4 Study IV

This study was born out of the ideas and results of the three previous studies. Although MAP showed great improvements (Study I), the rank-wise inspection measured by CG (Study II) gave somewhat disappointing results regarding the positive effect of RF. Statistically significant improvements were observed but

[82] Conventional relevance weights exclude explicit negative values. Therefore, even at very high ranks the performance curves keep going upwards. This may be considered counter-intuitive from the point of view of a user.

[83] See Figures 4-5 in Study III.

intuitively speaking the improvements seem to be of a rather modest size. Also, one feedback round was not enough to radically improve the result retrieved.

The results of Study II suggest that no feedback may be available in small browsing windows inspected, especially if high quality feedback is demanded. Yet the user may not be motivated to continue browsing the "failed" list of documents retrieved (to find feedback). In fact, the need for query modification may be most pressing in such situations where the initial query failed.

Faced with such concerns we decided in Study IV to model *direct* query modifications, instead of trying to find positive RF based on the initial search in vain. The idea to use direct query modifications was inspired by Järvelin et al. (2008).[84] Also the concepts discussed in Study III had an impact on Study IV: first, we conceptualize the effectiveness in sessions in terms of *success* or *failure,* and secondly, we acknowledge the "negativity" of users as an important aspect by focusing on *impatient* users - who prefer *short* queries (in most cases only 1-3 words), tolerate *limited browsing* (10 documents per query), and quit after finding *one* relevant document. Importantly, we experiment with using up to *five* queries per one topical session, instead of using only *one* feedback round (Studies I and II). In the light of traditional test collection-based IR experiments, the retrieval scenario of Study IV is somewhat unconventional. However, compared to real life one may argue that it represents a conventional way to conduct searching.

The concept of interaction decisions toward query formulation (and reformulation); browsing patience; and the relevance levels (from Studies I and II) are continued here. The study utilizes query data collected from test persons while the effectiveness of three prototypical short query session strategies[85] and one baseline strategy were compared under the laboratory conditions using simulation.

*Research questions*

In this study our research question is:

1. How effective are sequences of short queries combined with impatient browsing, compared to using one long query and patient browsing?

[84] Yet differently, Järvelin et al. (2008), had real test persons performing simulated interactive search tasks, while we performed surface level interaction simulations in the laboratory without users.

[85] See Section 3.4.4.

*Methods*

In Study IV realism is addressed by the involvement of test persons to intellectually collect words which are systematically utilized in session simulations to construct short-query sequences.[86] In the simulations we control how the query sequences are formed and how long the result list is browsed within a session.

*Results*

Our main experimental result was that even when highly relevant documents were required for success, the simple strategies were successful. In more than a half of the topics after the user had tried out only very few queries (e.g., only three *individual* query words) a highly relevant document was retrieved.[87] We also considered query formulation, query launching and browsing costs together. If only one query per topic is assumed, short queries seem inferior, but they make sense *as sequences* if the user wants to minimize the number of search terms used and accepts taking chances with individual queries.

The binary success measure we used is also justified. In our simulation we required finding only one relevant document[88] (using two relevance thresholds) but this type of simulations could be performed by assuming other success criteria.

[86] This approach resembles the simulated work task situation described by Borlund (2000) yet in simulated work tasks test persons are involved during the various phases of the retrieval.
[87] See Table 2 in Study IV.
[88] Sakai (2006) argues that finding exactly one relevant document with a satisfactory relevance level and high precision is an important IR task.

# 6. Discussion and Conclusions

The starting point of our study was the observed discrepancy between real life information retrieval process and the implicit assumptions of the traditional test collection-based evaluation. We proposed using *user interaction simulations* to bridge the gap between the observed differences.

Considering the justifications given for the simulated approach given in Chapters 2 and 3 there must be good reasons why interaction simulation approach does not constitute the mainstream of the laboratory based IR. We suggest that reasons for this include the following:

- lack of tradition to simulate real life interaction processes in RF and SS using traditional test collections without users[89]
- the dominance of single-query batch experimentation
- "multi-problem" nature of real life simulations; we encountered simultaneously the problems of needing to (i) discover and justify real life behavior which we abstract in simulations;[90] (ii) implement the interaction processes; (iii) run the experiment using manageable and acceptable attribute value combinations (browsing lengths during RF, the maximum number of RF documents used, etc.); and (iv) address the evaluation of the results.
- lack of methodological tradition; issues like freezing and residual collections have been discussed in the literature but in graded relevance environment we encountered novel problems.[91]

[89] White et al. (2004) simulate searchers following various relevance paths between granular document representations, like title or top-ranking sentences, which can be explored in various orders. The information along these paths is used for implicit RF, using various term extraction models, to select expansion terms describing the information viewed by the searcher.

[90] Information retrieval is affected by the task, situation and user, including the user's experience regarding the task, topic searched, the search system, techniques; costs and effects; and his creativity. Obviously it is difficult to simulate such (possibly essential) user attributes related to interactive behavior.

[91] For example, assuming that graded RF is given, how should we evaluate the result when a different relevance threshold may be purposefully used in feedback and final evaluation.

We first characterized the complexity of real life search process to explicate the differences between real life searching and the traditional TCE approach, and to justify our simulations as valid. We presented a verbal description of a simplified interactive RF process together with a more formal description. Then we constructed simple scenarios for experimenting with simulated user relevance feedback. We paid special attention towards highly relevant document because real users often prefer finding the best documents (see, e.g., Sakai, 2006).[92]

Our initial intention in Study I was to measure how the quality and quantity of relevance feedback is related to search effectiveness when the user attempts to give RF by recognizing feedback within a limited-size browsing window. We also wanted to compare direct feedback (RF documents selected by the simulated user) to pseudo-relevance feedback (*all* top documents accepted as feedback). High quality feedback performed best when stringent evaluation criterion was used, while mixed quality RF performed best when the liberal evaluation criterion was used. Pseudo-RF also improved search results by each relevance level but it was not very competitive when the stringent relevance criterion was used.

In Study II we focused on the user point of view and performed full freezing of the initial results retrieved and "seen" by the simulated user, while measuring the effectiveness rank-wise based on gain cumulated. Also in this setting RF gave significant improvements compared to the baseline. However, our main conclusion was that *mixed-quality* RF made sense. If a small browsing window is used in collecting feedback, high quality RF may simply not be available. Even when a large browsing window was used, mixed quality RF was beneficial.

In Study III we focused on negative user sentiments. The concept *negative higher-order relevance* makes such sentiments visible and allows explaining user stopping behavior which is important in an interactive context. In the course of writing this paper the idea of considering topical queries from the point of view of *successful* versus *failed sessions* became apparent.

The last paper was born out of the ideas and results of all three previous studies. Although MAP showed great improvements in Study I, the rank-wise inspection measured by CG in Study II combined with full freezing gave somewhat disappointing results regarding the positive effect of RF. Statistically significant

[92] The utility of marginally relevant documents is questionable when they do not contain information adding to the topical description (Kekäläinen and Järvelin, 2002b, Scholer and Turpin, 2009).

improvements were observed and they were in the positive direction, but intuitively speaking the improvements seemed to be of a rather modest size. Moreover, no feedback may be available in short browsing windows inspected, especially if high quality feedback is demanded. A real user might avoid browsing the "failed" list of documents. In fact, the need for query modification may be most pressing in those situations where the initial query failed.

Therefore, in Study IV we modeled query modifications directly. We conceptualized the effectiveness of sessions in terms of success or failure. We acknowledged the existence of negative sentiments by focusing on very impatient users. Such users prefer short queries, tolerate only short browsing, and quit after finding one relevant document. We also raised the number of query attempts up to *five* queries per topical session. These features are justified because real searchers may have limited willingness to devote time for inventing search keys and browsing the retrieved result, while they do utilize the interaction possibility to the full. The results showed that such approach typically leads to good enough result. Our research results can be briefly summarized as follows (Table 2).

*Table 2.* Main results of Studies I-IV summarized

| | Study I | Study II | Study III | Study IV |
|---|---|---|---|---|
| **Main results** | 1. Simulated user RF can be effective at all relevance levels.<br>2. High quality RF performs best at stringent evaluation criterion; "mixed quality" RF performs best using liberal evaluation criterion<br>3. Pseudo-RF improved search results by each relevance level but was not very competitive when stringent criterion was used. | 1. User viewpoint justifies freezing and directly rank-based methods like cumulated gain-based evaluation<br>2. RF significantly improves effectiveness despite full freezing; high initial effort pays off if the user is patient during evaluation. Mixed quality RF makes sense. | 1. NHOR-based curves visualize performance towards failure or success.<br>2. Straightforward to operationalize using negative gain values.<br>3. Visualization of important user sentiments; allows explaining stopping behavior. | 1. Sequences of extremely short queries are surprisingly effective – even though the users did not interact with the list of the documents retrieved. |

Our present study is limited by the fact that we did not vary the attributes of query and index types, ranking algorithms, and RF methods – as is typically done in traditional TCE experiments. Instead, we varied in RF simulations the maximum browsing length, the maximum number of RF documents used, and the relevance level in accepting documents as feedback. In pseudo-RF, we varied the number of feedback documents used. In session strategy simulations we varied the session strategies used. As our simulations were based on an existing test collection, we used well-defined topics, topical relevance judgments, and we assumed document independence during evaluation. Differently, in real life the information needs may be ill-defined; the concept of relevance is dynamic and it has several manifestations; and the documents are not independent from each other. We also abstracted away the issues related to document structure and search interfaces and modeled a simplified situation during RF and SS where the simulated user was assumed to inspect all documents - and always the whole documents - from rank 1 to N. However, in real life some interactions may be more likely than others due to, e.g., interface issues. For example, the user might be guided by misleading document summaries and not inspect all documents entirely (Turpin et al., 2009). Moreover, we assumed in RF simulations that the relevance level of the feedback document was estimated correctly by the simulated user. However, in real life the user might make errors and give (at least partially) incorrect relevance feedback. More fine-grained models can be developed in the future to include such RF issues into simulations.

Figuring out good query keys and combinations of keys may be a more important problem than ranking well for any query (Järvelin et al., 2008). The single query approach does not help us to identify which *moves* are effective between differently behaving queries.[93] Query and browsing-based simulations can be used to study this issue. Most query reformulations continue to be manual although people use available terminological support for query expansion. Because rapid, direct intellectual query reformulations seem to be an attractive option for real end users, studying such interaction via simulations is justified. It would be possible to study, e.g., the effectiveness of sequences of short queries assuming that they are implicitly structured (see Ruthven, 2008).

The SS simulations could be constructed based on the types of query modification strategies popular in real life in various stages of search tasks (see, e.g., Vakkari, 2001), extending the types of idealized strategies used in Study IV. Moreover, the effectiveness of query modification approaches suggested in the

[93] We may see the effectiveness of a system or a search approach in a new light if we change the view of the process modeled and the evaluation. Azzopardi (2009a) argues that by focusing evaluation on the query, as opposed to the system, a number of interesting research questions arise; e.g., how to model how users generate queries; how much effort should be spent querying; and what is the relationship between query effort and retrieval effectiveness.

online searching literature, e.g., “pair-wise facets” and “briefsearches” (Efthimiadis, 1996) could be explored via simulations in relation to a given success criterion. The effectiveness of various interactive query and browsing strategies could be studied even if they are not currently popular in real life.

To support such simulations *extended test collections* can be developed in the future. Such collections would cover the possible “terminological worlds” reasonably available for the simulated searcher during a session. To construct extended collections the main concepts and the conceptual relationships in the topics are analyzed, and their expressions in the (relevant) documents are observed by test persons, e.g., performing a simulated work task. Such collections would facilitate systematic construction of various query sequences (see Keskustalo and Järvelin, 2010a; Keskustalo and Järvelin, 2010b; Keskustalo et al., 2010).

Regarding the evaluation aspect it is important to notice that particular metrics used in a batch experiment may not reflect the user task – e.g., assuming precision-based user tasks, metrics like MAP containing a recall component may be meaningless in the user domain (Scholer and Turpin, 2009). Rank-wise inspection (combined with full freezing) based on, e.g., CG-based metrics is well suited for user simulations because it directly adopts the user point of view (see Järvelin et al., 2008).

Our simulations of user interaction aimed at extending the traditional laboratory view of IR by modeling various “what if” scenarios. Our models were justified by the user behavior observed in real life. In the future test persons could be involved to empirically validate (*a posteriori*) the extent to which a particular simulation is an accurate representation of the real world. Future research should also explicate interactive user behavior and users’ success criteria regarding both the retrieved result and their preferences for searching action – how to reach their goal. We have demonstrated such formative simulations based on a traditional test collection. Because valid evaluation must take into account the kind of behavior taking place in real usage situations, we expect that test collection-based interactive user simulations will become a popular way to perform experiments in the future.

# References

Aalbersberg, I. J. (1992) Incremental relevance feedback. In: Belkin, N. J., Ingwersen, P. and Mark Pejtersen, A., eds., Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, pp. 11-22.

Adams, E. & Rollings, A. (2007) Fundamentals of Game Design, Prentice Hall, New Jersey. 669 p.

Ahlgren, P. (2004) The effect of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database. Dissertation. Valfrid, Sweden, 2004. 166 p.

Azzopardi, L. (2007) Position Paper: Towards Evaluating the User Experience of Interactive Information Access Systems. In: SIGIR'07 Web Information-Seeking and Interaction Workshop, 5 p.

Azzopardi, L. (2009a) Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 556-563.

Azzopardi, L. (2009b) Usage Based Effectiveness Measures: Monitoring Application Performance in Information Retrieval. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 631-640.

Azzopardi, L., Järvelin, K., Kamps, J. & Smucker, M. D. (2010) Simulated Evaluation of Interactive Information Retrieval. SIGIR Workshop Proposal. 3 p.

Bates, M. J. (1989) The Design of Browsing and Berrypicking Techniques for the Online Search Interface. http://www.gseis.ucla.edu/faculty/bates/berrypicking.html (visited 7 October 2010).

Beaulieu, M. (2000) Interaction in Information Searching and Retrieval. Journal of Documentation, 56 (4), pp. 431-439.

Belkin, N. J., Cool, C., Croft, W. B. & Callan, J. P. (1993) The Effect of Multiple Query Representations on Information Retrieval System Performance. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 339-346.

Belkin, N. J., Cool, C., Koenemann, J., Ng, K. B. & Park, S. (1995) Using Relevance Feedback and Ranking in Interactive Searching. TREC 1995. http://citeseer.ist.psu.edu/belkin96using.html (visited August 14th, 2007).

Belkin, N. J. (1980) Anomalous States of Knowledge as a Basis for Information Retrieval. Canadian Journal of Information and Library Science, 5, pp. 133-143.

Birta, L. G. & Arbez, G. (2007) Modelling and Simulation: Exploring Dynamic System Behavior. Springer-Verlag, London, 2007. 454 p.

Blair, D. C. (1984) The Data-Document Distinction in Information Retrieval. Communications of the ACM, 27 (4), pp. 369-374.

Bookstein, A. (1983) Information retrieval: a sequential learning process, Journal of the American Society for Information Science, 34 (5), pp. 331-342.

Borlund, P. (2000) Experimental Components for the Evaluation of Interactive Information Retrieval Systems. Journal of Documentation, 50 (1), pp. 71-90.

Broglio, J., Callan, J. P. & Croft, W. B. (1994) INQUERY system overview. In: Proceedings of the TIPSTER text program (Phase I), pp. 47-67.

Chang, Y. K. , Cirillo, C. & Razon, J. (1971) Evaluation of feedback retrieval using modified freezing, residual collection, and test and control groups. In: Salton, G., ed., The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, London, 1971. pp. 355-370.

Cleverdon, C. W. (1991) The Significance of the Cranfield Tests on Index Languages. In: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and

Development in Information Retrieval, 3-12. http://delivery.acm.org/ (visited 7 October 2010).

Cleverdon, C. W., Mills, L. & Keen, M. (1966) Factors determining the performance of indexing systems, vol. 1 - design. Cranfield: Aslib Cranfield Research Project.

Conover, W. J. (1999) Practical Nonparametric Statistics, 3rd edition. New York: John Wiley and Sons. 584 p.

Cooper, W. S. (1973) On Selecting a Measure of Retrieval Effectiveness. Part 1. Journal of the American Society for Information Science, 24 (2), pp. 87-100.

Cosijn, E. & Ingwersen, P. (2000) Dimensions of Relevance. Information Processing and Management, 36 (4), pp. 533-550.

Dennis, S., McArthur, R. & Bruza, P. D. (1998). Searching the World Wide Web made easy? The cognitive load imposed by query refinement mechanisms. In: Proceedings of the 3rd Australian Document Computing Conference, Sydney, Australia. Sydney: University of Sydney, Department of Computer Science, TR-518, pp. 65-71.

Efthimiadis, E. N. (1996) Query expansion. In: Williams ME, ed., Annual Review of Information Science and Technology, vol. 31 (ARIST 31). Medford, NJ: Learned Information for the American Society for Information Science, pp. 121-187. http://faculty.washington.edu/efthimis/pubs/Pubs/qe-arist/QE-arist.html (visited 7 October 2010).

Fidel, R. (1985) Moves in online searching. Online Review, 9 (1), pp. 62-74.

Fuhr, N., Belkin, N., Jose, J. M. & van Rijsbergen, C. J. (2009) Workshop Report: Seminar 09101 – Interactive Information Retrieval. Dagstuhl, Germany. March 2009. 5 p.

Harman, D. (1988) Towards Interactive Query Expansion. In: Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 321-331.

Hersh, W. (1994) Relevance and Retrieval Evaluation: Perspectives from Medicine. Journal of the American Society for Information Science, 45 (3), pp. 201-206.

Hersh, W. & Over, P. (2001) TREC-9 Interactive Track Report. Retrieval Group Information Access Division, NIST, Gaithersburg, USA. May 14, 2001. 9 pages.

Hull, D. (1993) Using Statistical Testing in the Evaluation of Retrieval Experiments. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 329-338.

Jansen, M. B. J., Spink, A. & Saracevic, T. (2000) Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web, Information Processing & Management 36 (2), pp. 207-227.

Joachims, T., Granka, L., Pan, B., Hembrooke, H. & Gay, G. (2005) Accurately Interpreting Clickthrough Data as Implicit Feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 154-161.

Jordan, C., Watters, C. & Gao, Q. (2006) Using Controlled Query Generation To Evaluate Blind Relevance Feedback Algorithms. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL'06), pp. 286-295.

Järvelin, K. (2009) Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments. In: Proceedings of the 18th ACM Conference on Information and knowledge management, Hong Kong, China, 2-6 November 2009, pp. 2053-2056.

Järvelin, K. & Kekäläinen, J. (2002) Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS), 20 (4), pp. 422-446.

Järvelin, K. & Kekäläinen, J. (2000) IR evaluation methods for retrieving highly relevant documents. In: Belkin, N. J., Ingwersen, P., Leong, M-K., eds., Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 41-48.

Järvelin, K., Price, S. L., Delcambre, L. M. L. & Nielsen, M. L. (2008) Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In: Proceedings of the 30th European Conference in IR Research (ECIR'08), pp. 4-15.

Kando N. (2000) What Shall We Evaluate? - Preliminary Discussion for the NTCIR Patent IR Challenge (PIC) Based on the Brainstorming with the Specialized Intermediaries in Patent Searching and patent Attorneys. In: Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval. http://research.nii.ac.jp/ntcir/sigir2000ws/sigirprws-kando.pdf (visited 8 October 2010).

Kekäläinen J. (1999) The Effects of Query Compexity, Expansion and Structure on Retrieval Performance in Probalistic Text Retrieval. Doctoral thesis. Tampere, Finland: University of Tampere, Department of Information Studies. Acta Universitatis Tamperensis 678. 170 p.

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations – Comparison of the effects on ranking of IR systems. Information Processing & Management, 41 (5), pp. 1019-1033.

Kekäläinen, J. & Järvelin, K. (2002a) Evaluating Information Retrieval Systems under the Challenges of Interaction and Multidimensional Dynamic Relevance. In: Proceedings of the 4th CoLIS Conference, pp. 253-270.

Kekäläinen, J. & Järvelin, K. (2002b) Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology, 53(13), pp. 1120-1129.

Keskustalo, H. & Järvelin, K. (2010a) Query and Browsing-Based Interaction Simulation in Test Collections. In: Azzopardi, L., Järvelin, K., Kamps, J., Smucker M. D., eds., Proceedings of the SIGIR 2010 Workshop on the Simulation of Interaction: Automated Evaluation of Interactive IR, Geneva, Switzerland, July 23, 2010, pp. 29-30.

Keskustalo, H. & Järvelin, K. (2010b) Simulations as a Means to Address Some Limitations of Laboratory-based IR Evaluation. In: Larsen, B., Schneider, J. W., Åström, F., Schlemmer, B., eds., The Janus Faced Scholar: A Festschrift in Honour of Peter Ingwersen, Informationsvidenskabelige Akademi, Copenhagen, 2010, pp. 69-86.

Keskustalo, H., Järvelin, K. & Pirkola, A. (2010) Graph-Based Query Session Exploration Based on Facet Analysis. In: Azzopardi, L., Järvelin, K., Kamps, J., Smucker M. D., eds., Proceedings of the SIGIR 2010 Workshop on the Simulation of Interaction: Automated Evaluation of Interactive IR, Geneva, Switzerland, July 23, 2010, pp. 15-16.

Keskustalo, H., Järvelin, K. & Pirkola, A. (2006) The Effects of Relevance Feedback Quality and Quantity in Interactive Relevance Feedback: A Simulation Based on User Modeling. In: Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsirika, T., Yavlinsky, A., eds., Proceedings of the 28th European Conference on IR Research (ECIR’06), London, UK, pp. 191-204.

Keskustalo, H., Järvelin, K. & Pirkola, A. (2008a) Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value. Information Retrieval, 11 (3), pp. 209-228.

Keskustalo, H., Järvelin, K., Pirkola, A. & Kekäläinen, J. (2008b) Intuition-Supporting Visualization of User’s Performance Based on Explicit Negative Higher-Order Relevance. In: Myaeng, S-H., Oard, D. W., Sebastiani, F., Chua, T-S., Leong M-K., eds., Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 675-681.

Keskustalo, H., Järvelin, K., Pirkola, A., Sharma, T. & Lykke, M. (2009) Test Collection-Based IR Evaluation Needs Extension Toward Sessions – A Case of Extremely Short Queries. In: Lee, G. G., Song, D., Lin, C-Y., Aizawa, A. N., Kuriyama, K., Yoshioka, M., Sakai, T., eds., Proceedings of the 5th Asia Information Retrieval Symposium (AIRS´09), pp. 63-74.

Korfhage, R. R. (1997) Information Storage and Retrieval, Wiley, New York, 1997, 349 p.

Kuhlthau, C. C. (1991) Inside the Search Process. Journal of the American Society for Information Science, 42 (5), pp. 361-371.

Marchionini, G., Dwiggins, S., Katz, A. & Lin, X. (1993) Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. Library and Information Science Research, 15 (1), pp. 35-70.

Pirkola A., Leppänen, E. & Järvelin, K. (2002) The RATF Formula (Kwok's Formula): exploiting average term frequency in cross-language retrieval. Information Research, 7(2). http://informationr.net/ ir/7-2/paper127.html (visited 7 October 2010).

Pollock, S. M. (1968) Measures for the Comparison of Information Retrieval Systems. American Documentation, October 1968, pp. 387-397.

Price, S. L., Nielsen, M. L., Delcambre, L. M. L. & Vedsted, P. (2007) Semantic Components Enhance Retrieval of Domain-Specific Documents. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, Lisbon, Portugal, pp. 429-438.

Rocchio, J. J., Jr (1971a) Evaluation viewpoints in document retrieval. In: Salton, G., ed., The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, London, 1971. pp. 68-73.

Rocchio, J. J., Jr (1971b) Relevance feedback in information retrieval. In: Salton, G., ed., The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, London, 1971. pp. 313-323.

Ruthven, I. (2008) Interactive Information Retrieval. In: Annual Review of Information Science and Technology, vol. 42, 2008. pp. 43-91.

Ruthven, I. & Lalmas M. (2003) A survey on the use of relevance feedback for information access systems. Knowledge Engineering Review, 18(2): pp. 95-145.

Sakai, T. (2006) Give Me Just One Highly Relevant Document: P-Measure. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 695-696.

Salton, G. (1989) Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Reading, MA: Addison-Wesley. 530 p.

Sanderson, M. (2008) Ambiguous Queries: Test Collections Need More Sense. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 499-506.

Saracevic, T. (1975) Relevance: A Review of and a Framework for Thinking on the Notion in Information Science. Journal of the American Society for Information Science, 26 (6), pp. 321-343.

Saracevic, T. (1996a) Modeling interaction in information retrieval (IR): a review and proposal. Proceedings of the American Society for Information Science, 33, pp. 3-9.

Saracevic, T. (1996b) Relevance Reconsidered '96. In: Proceedings of the 2nd CoLIS Conference, pp. 201-218.

Saracevic, T. (2006) Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II. Advances in Librarianship, vol. 30, 2006. pp. 3-71.

Scholer, F. & Turpin, A. (2009) Metric and Relevance Mismatch in Retrieval Evaluation. In: In: Lee, G. G., Song, D., Lin, C-Y., Aizawa, A. N., Kuriyama, K., Yoshioka, M., Sakai, T., eds., Proceedings of the 5th Asia Information Retrieval Symposium (AIRS´09), pp. 50-62.

Siegel, S. & Castellan, N. J. (1988) Nonparametric Statistics for the Behavioral Sciences. Second edition. McGraw-Hill, New York.

Smith, C. L. & Kantor, P. B. (2008) User Adaptation: Good Results from Poor Systems. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 147-154.

Sormunen, E. (2002) Liberal relevance criteria of TREC – Counting on negligible documents? In: Beaulieu, M., Baeza-Yates, R., Myaeng, S. H., Järvelin, K., eds., Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 320-330.

Sormunen, E., Kekäläinen, J., Koivisto, J. & Järvelin, K. (2001) Document Text Characteristics Affect Ranking of the Most Relevant Documents by Expanded Structured Queries. Journal of Documentation, 57 (3), pp. 358-374.

Spink, A. (1997) Study of Interactive Feedback during Mediated Information Retrieval. Journal of the American Society for Infirmation Science, 48 (5), pp. 382-394.

Spink, A., Greisdorf, H. & Bateman, J. (1998) From highly relevant to not relevant: examining different regions of relevance. Information Processing & Management, 34 (5), pp. 599-621.

Spink, A. & Saracevic, T. (1998) Interaction in Information Retrieval: Selection and Effectiveness of Search Terms. Journal of the American Society for Information Science, 48 (8), pp. 741-761.

Stenmark, D. (2008) Identifying Clusters of User Behavior in Intranet Search Engine Log Files. Journal of the American Society for Information Science and Technology, 59 (14), pp. 2232-2243.

Swanson, D. (1977) Information Retrieval as a Trial-and-Error Process. Library Quarterly, 47 (2), pp. 128-148.

Turpin, A. & Hersh, W. (2001) Why Batch and User Evaluations Do Not Give the Same Results. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 225-231.

Turpin, A. & Scholer, F. (2006) User Performance versus Precision Measures for Simple Search Tasks. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 11-18.

Turpin, A., Scholer, F., Järvelin, K., Wu, M. & Culpepper J. S. (2009) Including summaries in system evaluation. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 508-515.

Vakkari, P. (2000) Cognition and changes of search terms and tactics during task performance: a longitudinal study. Proceedings of the RIAO 2000 Conference, Paris: C.I.D., pp. 894-907.

Vakkari, P. (2001) A theory of the task-based information retrieval process: a summary and generalization of a longitudinal study. Journal of Documentation, 57 (1), pp. 44-60.

Vakkari, P. (2002) Subject Knowledge, Source of Terms, and Term Selection in Query Expansion: An Analytical Study. In: Crestani, F., Girolami, M., Rijsbergen, C.J., eds., Proceedings of the 24th European Colloquium on IR Research (ECIR'02), LNCS 2291, pp. 110-123.

Vakkari, P. & Sormunen, E. (2004) The Influence of Relevance Levels on the Effectiveness of Interactive Retrieval. Journal of the American Society for Information Science and Technology, 55 (11), pp. 963-969.

Voorhees, E. & Harman, D. (2005) TREC: experiment and evaluation in information retrieval. MIT Press.

Voorhees, E.M. (2001) Evaluation by Highly Relevant Documents. In: Croft, W. B., Harper, D. J., Kraft, D. H., Zobel, J., eds., Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA, pp. 74-82.

Voorhees, E.M. (2007) TREC: Continuing Information Retrieval's Tradition of Experimentation. Communications of the ACM, 50 (11), 51-54. http://portal.acm.org.

White, R. W., Jose, J. M., van Rijsbergen, C. J. & Ruthven, I. (2004) A Simulated Study of Implicit Feedback Models. In: McDonald, S., Tait, J., eds., Proceedings of the 26th European Conference on IR Research (ECIR'04), Sunderland, UK, pp. 311-326.

White, R. W., Ruthven, I., Jose, J. M. & van Rijsbergen, C. J. (2005) Evaluating Implicit Feedback Models Using Searcher Simulations. ACM Transactions on Information Systems (TOIS), 23 (3), pp. 325-361.

Yang, Y., Lad, A., Lao, N., Harpale, A., Kisiel, B. & Rogati, M. (2007) Utility-based Information Distillation Over Temporally Sequenced Documents. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 31-38.

# Errata

The “Average Search Length” column in Tables 2-4 in Paper I contains errors. The correct figures, from top to bottom, are as follows:

Table 2: 30.0, 29.6, 27.2, 14.3, 10.0, 9.8, 6.3, 5.0, 3.5, 1.0.

Table 3: 30.0, 27.4, 21.1, 6.4, 10.0, 9.1, 3.6, 5.0, 2.4, 1.0.

Table 4: 30.0, 24.7, 14.8, 4.4, 10.0, 8.7, 2.5, 5.0, 1.9, 1.0.