FEZA BASKAYA

Simulating Search Sessions in Interactive Information Retrieval Evaluation

ACADEMIC DISSERTATION
To be presented, with the permission of
the Board of the School of Information Sciences
of the University of Tampere,
for public discussion in the Auditorium Pinni B 3116,
Kanslerinrinne 1, Tampere, on June 13th, 2014, at 12 o'clock.

UNIVERSITY OF TAMPERE

FEZA BASKAYA

Simulating Search Sessions in Interactive Information Retrieval Evaluation

Acta Universitatis Tamperensis 1949
Tampere University Press
Tampere 2014

UNIVERSITY
OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere
School of Information Sciences
Finland

The originality of this thesis has been checked using the Turnitin OriginalityCheck service in accordance with the quality management system of the University of Tampere.

Cover design by
Mikko Reinikka

Distributor:
kirjamyynti@juvenes.fi
http://granum.uta.fi

Acta Universitatis Tamperensis 1949
ISBN 978-951-44-9498-7 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1434
ISBN 978-951-44-9499-4 (pdf)
ISSN 1456-954X
http://tampub.uta.fi

Suomen Yliopistopaino Oy – Juvenes Print
Tampere 2014

YMPÄRISTÖMERKKI
MILJÖMÄRKT
441 729
Painotuote

Acknowledgments

Sitting in a time-machine called Earth, one does not even notice that time flies like an arrow. At the School of Information Sciences (SIS) which by the way means "fog" in Turkish, I started as a software developer 10 years ago. Having written Information Retrieval related software for some time I became a university student again, this time, a Ph.D. student. After trying to find the correct path in a foggy environment, I have seen the lights of a FIRE (Finnish Information Retrieval Experts), around which master chefs have gathered. Enthralled by the flickering fire, I got thirsty for more knowledge. Countless discussions around fire ushered the way to cooking this thesis. Finland's wonderful nature, with lakes, forests, and beauties gave inspiration for a myriad of recipes. In addition, the ingredients for the recipe are collected from the endless World Wide Web. Finally, the dish is now cooked with the kind advice of the chefs around the FIRE; one last challenge to stand is opponent's wisdom. Then, my precious Ph.D. thesis is ready to be served to the wide world. Hopefully, it instigates further recipes in this specific domain pursuant to Zeitgeist. Bon appetite!

First of all, I would like to thank my supervisors and co-authors Prof. Kalervo Järvelin and Dr. Heikki Keskustalo. It was a pleasure to work with you. Without your positive attitude, this piece of art would not have been created at all. I learned a lot during our weekly discussions. As Kal always said, I have a Bosporus bull attitude. I would like to have simple straightforward novel and palpable methods and the best results. However, I learnt how everything is relative, how differently one could interpret the facts, among other things. Moreover, Heikki's strictness about every detail just contributed to the perfection of the papers and this thesis. I am delighted in having both of you as my supervisors.

Further, I would like to thank my colleagues Dr. Sanna Kumpulainen, Dr. Paavo Arvola, Dr. Sami Serola, and Dr. Eija Airio for their nice discussion about moving heaven and earth as well as other Ph.D. students and FIRE group experts for supportive comments during FIRE seminars. I am grateful to Professors Jaana

Kekäläinen, Eero Sormunen, Reijo Savolainen, and Pertti Vakkari for providing knowledge and wisdom.

Furthermore, I am grateful to the pre-reviewers of my thesis, Dr. Leif Azzopardi, University of Glasgow, UK, and Assistant Professor Mark D. Smucker, University of Waterloo, Canada, for their assessments.

Finally, I want to thank all my friends and my family. I thank my parents, Sevim and Kemal, for raising us, and my brothers, Dr. Zafer Baskaya and candidate Dr. Alparslan Baskaya, for creating a competitive environment. Last, I owe my wife, Katri, a debt of gratitude for her constructive positive outlook on life, and our children, Sema Sofia and Martti Fatih for their very being and sweetness.

In case I forgot your name, I thank all of you who feel thanks are deserved.

Love, Peace, and Happiness!

Tampere May 9th 2014

Feza Baskaya

Abstract

Modern knowledge society would not be possible without Information Retrieval (IR), because of the ever-growing amount of information available on the Internet. Information Retrieval provides crucial ways of finding the proverbial needle in a haystack. While computing technology is nowadays ubiquitous, users interact with various computer interfaces with varying goals and time constraints in order to complete their tasks, which may be initiated by their work or leisure-related activity. Thereby, Interactive Information Retrieval (IIR), which is the subject of this thesis, constitutes an important part of task performance. Users' interaction is shaped by users' personal and search characteristics, such as query formulation strategy, strategies for scanning and assessing of the search results, as well as users' feedback behavior.

Experimental evaluation is essential to the assessment of the effectiveness of IR systems. The traditional approach to measuring the effectiveness of diverse IR systems goes back to the Cranfield tests in the 1960s. However, neither user characteristics nor time are considered in the traditional evaluation process. In the Cranfield-type tests, still popular today, users are taken into account only marginally and their interests are represented in relevance assessments, evaluation metrics and topics to some extent. However, interaction with an IR system can be dissected more precisely and users' interaction during a search session can be divided further into subtasks. This in turn affects the evaluation process of IR systems. Moreover, users' feedback during a search session, which may be of high or poor quality, can be exploited to improve the search results. This again influences the effectiveness of search systems. In the present thesis, we examine the effects of users' characteristics and the relevance feedback behavior on search effectiveness. While conducting interactive experiments with test persons is costly in terms of time and resources, our experiments are based on user behavior simulations, which can be conducted within a short time, even though a vast number of sessions representing various user characteristics are reproduced.

Our study suggests that relevance feedback can be utilized in conjunction with classification algorithms to improve search results. Further, a realistic level of fallibility in the feedback process does not deteriorate the search outcomes significantly. When time is taken into account, it plays a major role in the evaluation process. Comparing the different search environments and strategies may be considered in respect to time expended during the search session. In that case, traditional evaluation metrics may deliver misleading conclusions in experiments. Further, we examined all possible query formulation strategies. Our experiments indicate that there is no single winning strategy that performs best across all topical search tasks. Moreover, we show that conventional IR experiments are not aware of user-dependent search variables such as query formulation, result scanning and assessing behavior, which govern the subtasks of the search process and the effectiveness of IIR. Therefore, the effect of these variables should be taken into account in the IIR evaluation process.

Finally, this thesis contributes to the methods of interactive information retrieval by better regarding the real life context and by simulating users' characteristics in information retrieval test environments. Consequently, the more users' behavior is perceived, recognized and understood, the more user friendly and effective information retrieval systems may be constructed.

Table_of_Contents

List of original publications

This thesis consists of a summary and the following original research publications, reprinted here by permission of the publishers.

I. Baskaya, F., Keskustalo, H., & Järvelin, K. (2013a). Effectiveness of search result classification based on relevance feedback. *Journal of Information Science, 39*(6), 764-772. doi:10.1177/0165551513488317

II. Baskaya, F., Keskustalo, H., & Järvelin, K. (2011). Simulating simple and fallible relevance feedback. In: *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, (pp. 593-604), ECIR 2011, Springer. Berlin Heidelberg. doi:10.1007/978-3-642-20161-5_59

III. Baskaya, F., Keskustalo, H., & Järvelin, K. (2012). Time drives interaction: Simulating sessions in diverse searching environments. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 105-114), SIGIR 2012, ACM. Portland, Oregon, USA. 105-114. doi:10.1145/2348283.2348301

IV. Baskaya, F., Keskustalo, H., & Järvelin, K. (2013b). Modeling behavioral factors in interactive information retrieval. In: *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, (pp. 2297-2302), CIKM 2013, ACM. San Francisco, California, USA. doi:10.1145/2505515.2505660

These publications will be referred to as Studies I-IV in the summary part of the thesis.

Research contribution of the author

First of all, this work and the present author stand on the shoulders of giants, who pioneered the way by creating the wonderful domain of Information Retrieval. Without the invaluable contributions of the supervisors, this thesis would not have seen the light of day.

In Study I, the present author created the research questions; collected and analyzed the data, designed and developed the necessary programs, implemented them, and then analyzed, compared and evaluated the research results, prepared the results for the publications, and wrote the articles with the help of co-authors.

In Study II, III, and IV, the present author contributed to the creation of research questions, collected and analyzed the data, designed and developed the necessary programs, implemented them, and then analyzed, compared, and evaluated the research results, prepared the results for the publications, and wrote the articles with the help of co-authors.

Prof. K. Järvelin and Dr. H. Keskustalo acted as the supervisors of the present author.

1. Introduction

Information Retrieval (IR) (Manning et al., 2008; Ricardo, 1999) is indispensable to our modern knowledge-based society. Modern information environments are becoming large and complex as well as ubiquitous, because the amount of available heterogeneous information grows exponentially each year (Alpert & Hajaj, 2008). Almost every aspect of our lives and every profession are affected by the information available on the Internet.

In order to gain knowledge, information should be obtained and analyzed by information users. In the first place, users should explicate their information need in a predefined way for a certain information retrieval system. Many different information objects such as documents, news, tweets, pictures, videos, audios, maps and 3D structures require various information need representations, not to mention the representation of those information objects in computer systems. However, one well-established communication medium is based on natural languages, or more precisely on the representation of words. Not only can documents, news, tweets, etc. be represented as text, but also other types of information objects such as audio-visual elements like pictures, videos and music can be described with words, which alleviate the possible problems of representation and access of those information objects. Therefore, information retrieval based on textual documents plays a major role in the research community and in real life. Consequently, the present research focuses on text-based document retrieval.

The history of document retrieval goes back to library science, where the documents were cataloged and accessed via catalog cards (Ruthven & Kelly, 2011, pp. 1-14). The categorization of documents was carried out according to salient features like title, author, publishing date and limited number of content keywords. However, emerging computer systems paved the way for automatic indexing of the full content of documents. Having all the content indexed, users were able to access and search appropriate documents according to their information need through search engine user interfaces. At first, Cleverdon et al. (1966) set up an experimental environment for IR experiments, in which documents were indexed by content features and retrieved via queries, and then evaluated in a batch mode. This experimental setup is better known as Cranfield IR evaluation, sometimes also called the laboratory IR, which can also be described as system-centered/oriented IR. However, while system-oriented IR focuses on performance and effectiveness, designing a good IR system depends not only on system-oriented performance issues but also on understanding the users who interact with the system (Ingwersen & Järvelin, 2005, pp. 111-258).

Research and industry efforts in IR bifurcate into two areas; the first being system-oriented research and development, and the second a user-oriented, academic research field. Most of the effort in research and development is spent on system design, development and evaluation. Even though these evaluation efforts take the user into consideration by including predefined relevance judgments and diverse evaluation methods, they are limited in nature. As humans are diverse, so are IR system users. Accordingly, the interaction of the user with IR systems exhibits miscellaneous behavior, which is lacking in the design and implementation of the pertinent systems. On the other hand, conducting comprehensive user studies is not only intricate but also prohibitively expensive. Unsurprisingly, academic studies on user-oriented IR usually employ a small number of users in their studies. This in turn confines the expressive power of those studies in terms of the generalization of hypotheses claimed. To bridge between system-oriented and user-oriented IR, in the present thesis we simulate the user characteristics in respect of information retrieval interaction. Thus, we not only circumvent the peculiarities of the individual user characteristics, but also enable the system-oriented IR to respect the user behavior and improve the capability of IR systems to utilize an enormous number of simulated users in laboratory experiments.

Real life information retrieval takes place in sessions, where users search by iterating between different subtasks through an interactive interface (Marchionini, 1995, pp. 27-60). As an overly simplified view, after examining results, users either modify the initial query or supply relevance feedback (RF) (Ruthven & Lalmas, 2003), which means users give feedback to the search system by indicating the relevant documents from the result list and continue the session until the information goal is achieved or the session is abandoned because of frustration or lack of time. Thereby, questions arise such as how relevance feedback can be utilized in the search system, how RF affects the IR performance when fallible feedback is provided, where the limits of effectiveness of diverse interactive searching strategies in different searching environments under overall cost constraints are, what kind and how effective the optimal sessions are under varying goals and constraints, and human stochastic behavior.

Psychological and/or social aspects of user behavior can be simulated in experimental designs according to experiment design (Ruthven, Lalmas, & Van Rijsbergen, 2003). Germane aspects of users should first be characterized to be exploited in the simulation. Among others, user's relevance feedback, fallibility in user's feedback, user's behavior under time pressure, endurance and scanning strategies in result scanning, and query modifications strategies in sessions are some of the simulated aspects in the present thesis.

The present thesis focuses on the simulation of user behavior and its effects on IR evaluation, and aims to answer research questions related to relevance feedback and multi-query session evaluation.

In previous RF studies, RF has been used to learn better queries in order to improve search result rankings after user's feedback. Those studies utilize query expansion methods to create better queries, which are consequently executed by the retrieval system. Instead of query expansion methods, we are interested in applying classification algorithms to improve result rankings without executing any further expanded queries. Accordingly, in Study I the main research question is: given RF on the first result page, assuming ten document surrogates are shown to a simulated user, is it possible to learn effective classifiers for the following result pages? Furthermore, we query issues such as how this novel classification approach depends on initial query length and how the effectiveness of this approach depends on diverse classification methods and term space reduction algorithms, which attempt to sort out the insignificant document terms.

Traditional RF studies assume perfectly correct RF, which means users are required to identify relevant documents in the initial results. In Study II we challenged that point and exercised progressively less perfect RF. This was motivated by the user studies, which expose fallible user behavior during RF (Vakkari & Hakala, 2000). Consequently, in Study II the overall research question is: how does RF affect information retrieval performance when short initial queries, which are one to three words long suggested by real persons, are employed and fallible feedback, assuming that users may err when they indicate the relevant documents, is provided? Further, we are also interested in finding out whether mistakes in RF affect the quality (relevance level) of the documents found.

In real life, users interact with retrieval systems on different devices such as smartphones, desktop computers or tablets. Again, these devices lend themselves differently regarding user interaction. This in turn affects the time users spend in order to achieve search goals. However, time aspects of retrieval results have not been considered in commonly applied evaluation methods. Besides, some users are fond of having only highly relevant documents, while others would be perfectly satisfied even with marginally relevant documents. Consequently, varying search goals and time constraints encourage us to find out their effects on IR evaluation. Hence, in Study III we explore how various devices affect information retrieval sessions under overall time constraints and what the proper evaluation methodology is when time is taken into account. Moreover, we also explore all the search strategies which are query sequences applied during a search session, in order to find the best and worst sessions and compare them to query patterns frequently observed in real life.

Classical studies assume an average user, who interacts with a retrieval system in a predictable and regular way. However, users are diverse and not always predictable. Moreover, because of numerous reasons, they can make mistakes such as skipping relevant documents when examining a result list. Thus, in Study IV we analyze what kind and how effective the optimal search sessions are under varying search goals and time constraints, provided that both ideal and stochastic human behavior is regarded. In addition to the simulation variant in Study III, we further elaborate the search process with more detailed subtasks. With ideal human behavior we mean that users make no errors during the search process, or to be more precise, users scan all documents one after another, click every relevant document without making any judgment errors, read them and judge their relevance correctly. In contrast, fallible human behavior means that users may well err during the search process, in other words they may skip some relevant documents, read non-relevant ones, judge them as relevant or judge the relevant ones as non-relevant by mistake.

Besides, we are methodologically interested in the simulation of a behavioral model based on comprehensive session subtasks and fallible human behavior.

The rest of this thesis is organized as follows. Chapter 2 briefly introduces Information Retrieval (IR), while Chapter 3 addresses the simulation of Interactive IR (IIR). In Chapter 4 the evaluation issues in IIR are handled. The summaries of the contributed studies are presented in Chapter 5. Chapter 6 discusses the results, draws conclusions and proposes future research.

2. Information Retrieval

Human information behavior consists of phenomena such as information needs, information seeking, searching, browsing, finding, judging, usage, communication, sharing, transfer, management, information habit, and information style, which in brief means any information-related human behavior (Ruthven, 2008).

Human information behavior can be modeled in numerous ways with a focus on different aspects. The history of IR has witnessed many such models, which are still valid and consider various aspects from diverse point of views. Ruthven (2008) and Toms (2013) discuss some of those models. In general, these models lay out the information landscape which characterizes human information behavior.

On the other hand, interactive information retrieval models are formed to describe information retrieval interaction, which is the focus of this thesis. Among others, we can list some significant and salient ones like Belkin's anomalous state of knowledge (ASK) (Belkin, 1980), Ingwersen's cognitive model (Ingwersen & Järvelin, 2005), Saracevic's stratified model (Saracevic, 1997), and Bates' berry-picking model (Bates, 1989).

In this chapter, we first introduce traditional information retrieval, and then describe interactive information retrieval. The third section discusses relevance feedback, which can be categorized into explicit, implicit and pseudo-relevance feedback. Finally, we discuss some common classification methods that we applied for RF in the present thesis.

2.1 Traditional Information Retrieval

Information retrieval systems store and manage information items, e.g., text documents, as well as enable users to access them efficiently. With traditional Information Retrieval (IR) we mean system-oriented IR, which focuses on documents and document collections, matching algorithm(s) to retrieve relevant information items to stated queries, and relevance judgments about documents in relation to queries.

Figure 1 depicts the traditional IR process, which is also called the laboratory model of IR. Figure 1 is adapted from Ingwersen and Järvelin's (2005, p. 5) schematized system-oriented IR Model. The main focus of the system-oriented approach is the representation of documents and search requests as well as their matching process. The user's involvement is confined to relevance and possible feedback judgments. Moreover, the relevance judgments of documents were created once by persons who may be developers of the experimental environment. In this view of IR, documents are represented and stored in a database corresponding to the applied retrieval model. Thereafter, the user's information need is translated into a search request, which is in turn represented as a query for the matching process. However, neither the task, which causes the user's information need, nor the user's real information context is taken into account in any way. Nevertheless, the matching algorithms deliver more or less relevant documents according to the match between the presentations of documents and query. At this stage it is possible to exercise feedback and modify the query. Results can now be evaluated by comparing the output documents against the recall base via diverse evaluation measures (e.g., Demartini & Mizzaro, 2006; Su, 1992).

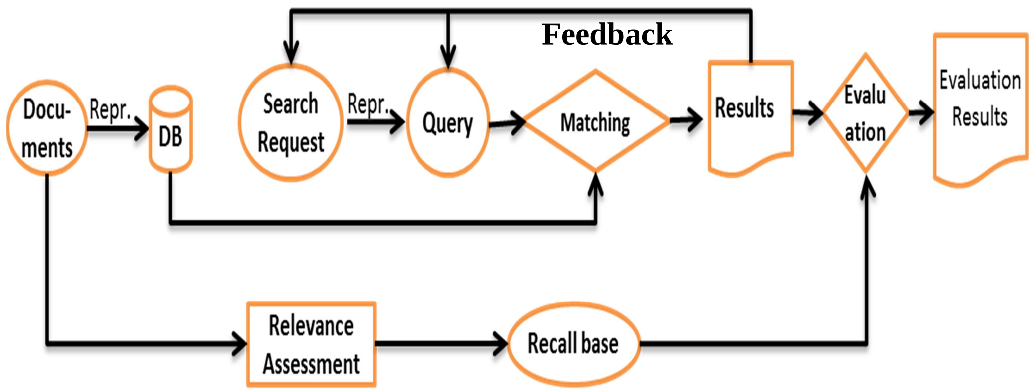

Figure 1. System-oriented view on IR (adapted from Ingwersen and Järvelin 2005, p. 5)

Retrieval effectiveness is dependent on the selected retrieval model. Belkin and Croft (1987) classify retrieval models into two main branches, namely exact and partial matching models. Exact matching models are developed on the basis of Boolean algebra. For this type of model, queries are meticulously constructed with the help of Boolean operators. Boolean logic-based retrieval models only deem those documents which exactly match the Boolean query as relevant, for example for the query "Information AND Retrieval", both keywords must appear in the relevant documents. Consequently, this model is robust but very strict, i.e., there is no possibility to obtain partially matching documents. Moreover, the delivered results list is not ordered according to relevance, though some ordering criteria like date or author's name can be enforced. Even if Boolean systems seem to be obsolete nowadays, still some specific domains like legal domain can require recall-oriented retrieval, which can be provided by Boolean systems.

On the other hand, in order to allow partially matching documents to be listed on the results list, partial matching models such as vector space models (VSM), probabilistic retrieval models and more recently probabilistic language models have been developed (Croft et al., 2010, pp. 233-296; Manning et al., 2008, pp. 109-134 & 219-252).

VSM was first realized in Salton's Smart information retrieval system (Salton, 1970). VSM represents both documents and queries as vectors in multidimensional space, whose dimensions consist of keywords. Every vector representing documents and queries can be built up with term weights like tf.idf (**t**erm **f**requency multiplied by **i**nverse **d**ocument **f**requency) (Belew, 2000, pp. 96-97) in respective documents and queries as axis values in each pertinent dimension. The similarity between a document and a query is calculated, for example, with the cosine similarity measure, which gauges the angle between two vectors. Then the documents can be ranked according to the cosine values in descending order. VSM is based on vector algebra, and is therefore mathematically founded, whereas its applicability in IR may be arguable from the justification point of view.

Probabilistic retrieval models (PRM) (Croft et al., 2010, pp. 233-296; Manning et al., 2008, pp. 219-236) are based on probability theories, especially the probability ranking principle, which means ranking by the decreasing probability of relevance of documents to a query. Documents can be ranked by the proportion of the probability of relevance and the probability of non-relevance ($\frac{P(R|D)}{P(NR|D)}$). PRM utilizes Bayes' rule for replacing the posterior probability P(R|D), the probability of relevance given to a certain document in the context of a current query, with the prior probability P(R) and the likelihood P(D|R). Applying Bayes' rule for both probabilities (e.g., P(R|D) and P(NR|D)) transforms the above proportion to $\frac{\mathrm{P(D|R)\cdot P(R)}}{\mathrm{P(D|NR)\cdot P(NR)}}$. Because the prior probability of relevance and non-relevance P(R) and P(NR) are the same for all documents, they are just playing a scaling factor for document scores, they can be removed from the formula.

Because the real probabilities are unknown, the probabilities in the formula above are estimated by diverse probability estimation methods in many different PRMs. For example, in the binary independence model (BIM), independence of the terms is assumed and term frequency in documents is taken into account simply as a binary feature. Hence, P(D|R) is estimated as a product of the probability of presence of a term and the probability of absence of a term in relevant documents. Given a query, the score for a document is the proportion (likelihood ratio) of the products of the term probabilities for all matching terms for relevance and non-relevance, which is usually converted to the sum of logarithms of term weights, because of mathematical precision concerns in computer memory systems. Because initially no relevant set is known, the pertinent probabilities are often set to a constant. Finally, when the proportions of the probabilities represent the term weights in a document, the similarity to the VSM model will be obvious.

Yet another probabilistic model introduced to the IR community is borrowed from language technologies. Language models (Manning et al., 2008, pp. 237-252) are applied in speech recognition, machine translation, spelling correction and other domains. Language modeling is based on probabilistic language models, which are estimated for every document in a collection. In order to rank the documents, document models are utilized to calculate the probability of generating the query. In language modeling, finite automata are exploited, for example, to generate the probabilities instead of generating strings for a language.

The probability of a query can be decomposed into the probability of each successive keyword conditioned on earlier keywords. Language models in IR are usually built from a single document. Therefore, there is not enough data to model complex conditional probabilities. The simplest possible language modeling, namely unigram language modeling (Manning et al., 2008, pp. 237-252), assumes the independence of terms; hence the probability formula is reduced to a probability calculation without a conditioning context. Assuming unigram language modeling, the probability of a query, especially in the query likelihood model, can simply be generated by multiplication of the probability of each query term with the help of the language modeling of the pertinent document. In case the query word is missing from the document language model, the probability of that query word will be zero, which requires special handling, or ‘smoothing’ (Zhai & Lafferty, 2001). Smoothing not only adds a fraction of probability to every word, but also discounts the non-zero probabilities. Having calculated these probabilities, documents can be ranked accordingly in descending order. As mentioned above, during implementation the multiplication of small numbers is replaced with a summation of logarithmic values; because the logarithmic function is a monotonic function, the ultimate order does not change, which is important for ranking the documents correctly.

In the current thesis we employed the Lemur/Indri search engine (Strohman et al., 2005) to conduct the experiments, because it is one of the very effective retrieval systems available as open source software. The Lemur/Indri search engine allows searching based on language modeling. Moreover, smoothing via Dirichlet priors (Zhai & Lafferty, 2001) is applied in order to avoid mathematical conundrums caused by the terms that do not exist in documents.

2.2 Interactive Information Retrieval

Ingwersen and Järvelin (2005, pp. 313-357) depicted an evaluation framework for interactive information retrieval. They tried to set IIR into various self-contained contexts, which resemble a Russian matryoshka doll. The framework starts from the socio-organizational and cultural context and ends at the IR context. Thereby, IR effectiveness can be evaluated in the socio-organizational context with socio-cognitive relevance, in other words, the quality of work task results. If the socio-organizational context is opened, the work task context appears. Again, IR effectiveness can be evaluated by the quality of information and work process or results. Under the work task context, the seeking context is located, in which IR effectiveness can be evaluated by the quality of the seeking process. The final context represents the IR context, in which the effectiveness of IR systems is traditionally evaluated by measures based on recall, precision, and efficiency, for example. Moreover, search engines, which are the core of the IR context, can be evaluated among others by the effort in using criteria such as comprehensibility of interface and query language, support in query formulation, and results presentation. In other words, these evaluation criteria need a user's subjective opinion. However, existing methods and measures for the evaluation of systems are user-agnostic. In the present thesis we try to close the gap between traditional IR and user studies by respecting the user characteristics such as his/her fallibility of feedback (i.e., pinpointing relevant information), frustration and context in time, space, and user's search goals (see Figure 2. Dimensions for extending IR) (Kamps et al., 2009).

In real life, users do not usually have ready-made topics or queries at their disposal. Instead, they are confronted with a problem or task, which is the source of the users' information need. This motivates them to seek information about the task or problem at hand. Thereby, different users cope with diverse information sources and with their access methods differently. Moreover, users' problems or tasks can be either work or leisure-related (Vakkari, 2003). This often sets some time and /or goal constraints. In addition, users have miscellaneous traits and backgrounds such as education, gender, domain knowledge, and perseverance which affect how they conduct the search process with an information retrieval system. In addition, the effect of numerous search interfaces of ubiquitous computer systems such as desktop computers, mobile phones, tablets and even smart television sets also have an influence on the search process. However, all these variables are not, at least directly, taken into account in the design and evaluation of interactive information retrieval systems. For example, system-oriented IR experiments, which hardly represent IIR experiments and systems, assume a user model representing ideal users, single query sessions, well-defined topics and queries, topical relevance, and document independence in order to design, develop and evaluate a user's information retrieval process.

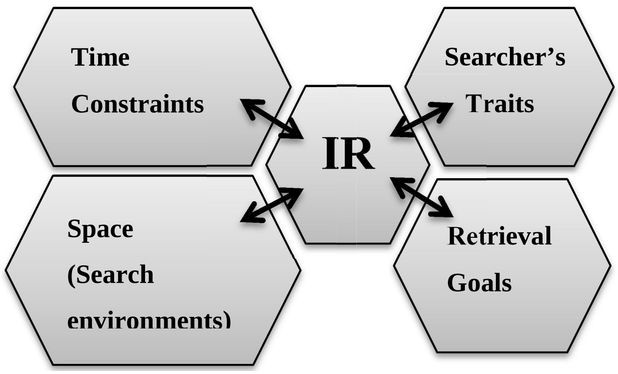

Figure 2. Dimensions for extending IR evaluation experiments

Keskustalo (2010) described six major limitations of traditional IR in his thesis:

1. No explicit user modeling
2. Single query sessions
3. Well-defined topics and queries
4. Topical relevance
5. Document independence
6. Challenge of traditional evaluation

The first limitation was about explicit user modeling, which means that traditional IR lacks an explicit user modeling which represents an individual user with a certain mental state, learning capability, educational background, and gender, inability in many respects, and a dynamic view of relevance (Kelly, 2009). However, taking all these user attributes into account would also complicate the evaluation process and even make it difficult to compare the results of different systems. Nevertheless, users in real life are different and exhibit different searching behaviors, and the systems should be evaluated accordingly. Therefore, we suggest explicit user modeling, which takes the personal traits and backgrounds of the users into account. In our studies we have incorporated users' search and query formulation strategy, recognizing the relevant snippets and documents, perseverance as frustration, and time and goal dependent behaviors, among other things.

The second limitation of traditional IR was about the single query sessions, which is justified from the batch processing point of view of system-oriented IR. However, users do not have all the query words initially at hand when they start to pose the query to a search system, let alone the search topic (Marchionini, 1993). Users learn while searching. Consequently, users often pose multiple queries during a search session. Another reason that users modify the query is the ambiguity of the query, e.g., in cases where homonyms are used. A search session ends when either the user's information need or goal is satisfied, or users give up because of frustration due to unsatisfactory results or a lack of time. Accordingly, multiple query sessions with several query modification or formulation strategies are simulated in our studies to explore the effects of multiple query sessions.

Third, well-defined topics and queries were a limitation mentioned by Keskustalo (2010). As a relic of Cranfield experimental design, topics and verbose queries usually constructed from topic description are quite common in IR experiments. Fixed information needs cast as topics and corresponding relevance judgments constitute a typical experimental setup for IR experiments. However, a user's information need is fuzzy at the beginning of a search session, and the information need may crystallize itself during the search process. Moreover, a user's relevance perception changes as they learn from inspected snippets and documents.

In addition to that, users in real life often prefer short queries (Jansen et al., 2000). In this thesis we also employ realistically short queries produced by real people for topics, which are predefined for certain information needs with corresponding relevance judgments in standard test collections. Thus, even though we do not circumvent all the limitations about topics, queries and recall bases, nevertheless we bring human-generated and realistically short queries into research settings.

The fourth limitation was about topical relevance, which can be described by the relationship between a topic expressed in a query and a topic covered by an information object (Saracevic, 2006). According to Saracevic's (1997) relevance system, relevance can be motivational, situational, cognitive, topical, and algorithmic. For example, situational relevance affected by e.g., time pressure, and motivational relevance affected by e.g., commitment to task, or cognitive relevance affected by e.g., domain knowledge and expertise can easily affect the information retrieval performance. Consequently, there is a need to deal with all these types of relevance in IR evaluation. Therefore, we simulate not only the topical relevance but also some aspects of situational, motivational and cognitive relevance through constraints, goals, and fallibility.

Document independence was the fifth limitation in Keskustalo's thesis (2010), which means relevance judgments in a recall base are based on individual documents and their mutual effects are omitted. However, a user's relevance perception changes with every inspected document or snippet. Moreover, recurrent information in a result list can affect users' relevance judgment (Kekäläinen & Järvelin, 2002). In spite of this we do not abandon document independence assumption, especially since the recall base which we utilize in our simulations is constructed under this premise. Still, we applied result list freezing (Keskustalo et al., 2008; Ruthven & Lalmas, 2003) in the evaluation process in order to alleviate the effect of recurrent documents. Besides, setting an appropriate goal like "find one relevant document" (Sakai, 2006) also contributes to the elusion of the independence limitation.

The challenges of traditional evaluation are presented as the last limitation in IR in Keskustalo's thesis (2010), which means inadequate evaluation metrics from the user's point of view are applied to measure the outcome of IR experiments. However, user's costs for example in terms of time expended and the frustration of the user with a futile system should be taken into account. Indeed, we will discuss the risks of traditional metrics when time is considered as a component in evaluation. Apart from this, we will present a formula to describe the perseverance behavior of searchers in our simulations.

In addition to the six limitations described by Keskustalo (2010), we discuss the following novel issues in the present thesis. We show the effects of various search interfaces during the search process (Kamvar et al., 2009). While traditional IR usually assume a typical search interface for experiments, we considered different types of search interfaces and their effects on the search process in respect of the costs involved and the utility gained. Thereby, instead of the utility a search session produces, we set time constraints and gain goals and investigated the best-performing patterns which are governed by user habits and behavior. Not only are the prototypical patterns investigated, which are prevalent for typical users as earlier research (Vakkari, 2000) showed; we also investigated a comprehensive set of search patterns in order to find any better strategies, which can surpass the common prototypical strategy outcomes.

Even though users are tacitly integrated into traditional IR research with relevance judgments, users' personal differences are not really regarded. For example, users' understanding of snippets (Turpin et al., 2009) or users' relevance perception can vary. This can lead to accepting various less relevant or even non-relevant documents as relevant. In our studies we examine various behavior-related variables such as clicking a snippet and judging a document, which are then represented by various probabilities. Another traditional assumption about users is that they are perfect. However, as we know, to err is human. We examine the fallibility of searchers[1] during RF and scanning search results. Yet another characteristic assumption about searchers in traditional IR is their robust perseverance during result inspection. In fact, users may get frustrated, especially when they encounter unproductive search results. Thereafter, users either abandon

[1] The terms *searcher* and *user* are interchangeably used in the present thesis.

the search session altogether or formulate another query. Indeed, this is an issue where we look forward to formulate frustration as a skipping probability. It represents a user's perseverance during a search session.

2.3 Relevance Feedback

Before even trying to explain relevance feedback (Harman, 1992; Ruthven & Lalmas, 2003), the following question should be answered: What is relevance? Let us first describe relevance in the context of IR. First of all, relevance can be classified into five sub-categories. The first of these is system relevance, which is related to the order of documents produced by the retrieval system and users' request; the second is topical relevance, which represents the topical relationship between documents and queries; the third is cognitive relevance, which is related to the mental level of receiving pertinent information; the fourth is situational relevance, which takes into account the situations e.g., the time pressure the user is subjected to and the effort they expend to carry out their tasks; and the fifth motivational relevance includes the user's frustration and lack of accomplishments (Saracevic, 1996). Under this classification of relevance we can now utilize some aspects of these relevance types that are appropriate to the present thesis.

Furthermore, relevance is not a constant – it changes during an information retrieval session, because the better the information need is understood by the user, the more precisely the relevance of documents can be judged. It is also reasonable to see this the other way round, in that the information needs of a user changes during a search session, and this in turn affects the relevance judgments (Saracevic, 2007; Vakkari & Hakala, 2000).

In order to improve information retrieval effectiveness, relevance feedback (RF) can be utilized (Ruthven & Lalmas, 2003). RF means that an information retrieval system utilizes feedback supplied by users either explicitly or implicitly inferred from user's behavior on a search result page in order to improve subsequent search results and their ranking. IR systems can exploit the given relevance feedback in various ways, for example a frequently employed method is the use of query modification (QM) (Carpineto & Romano, 2012; Efthimiadis, 1996). Certainly, RF implementation depends on the information models used, as different models require the integration of the RF in different ways into the model. However, a novel approach to applying RF is suggested in this thesis, which leverages the classification methods for RF by classifying the as-yet unseen results with the help of a classifier, which is trained on documents from RF.

There are two types of relevance feedback: explicit and implicit. Explicit RF requires explicit input from the retrieval system user, whereas implicit RF discovers RF information from the user's behavior and actions on the result page (White, 2011).

2.3.1 Explicit Relevance Feedback

Explicit RF requires explicit user feedback, which means users of IR systems should contribute the feedback information to the system in some form. However, there are many challenges to relevance feedback, starting with the cognitive load of the user. Depending on the user's mental capacity and current task difficulty, supplying RF can provide an extra burden on the users. Moreover, RF requires additional effort both from users and system developers. In spite of this, additional effort can be compensated with better ranking of the search results. Although RF is not usually realized as an essential part of a routine search process, additional effort can be justified with the reward of better search results. For example, Google offers a "similar" or "related articles" link, which appears from time to time either in the results snippet or in the results preview window, depending on the presented information object, for requesting relevant pages. Yet another difficulty could emerge with peculiar complex document types which users have to cope with. Complex documents, such as multi-topic or partially relevant documents, are those which can influence the outcome of the RF process. Such documents can accommodate both relevant and non-relevant keywords, so that they may cause query drifting in query modification realization of RF (White, 2011).

As mentioned briefly in the previous section, the nature of relevance judgments makes the relevance feedback harder. Another aspect of relevance is its partiality. As the majority of relevance assessments are based on binary decisions, a document is either relevant or non-relevant (Voorhees & Harman, 2000). However, if the complexity of the documents and information needs are considered, it can be very quickly understood that the binary approach is not sufficient for judging the relevance of documents.

Instead of binary judgments, Sormunen (2002) created a graded relevance-based collection, in order to study the effect of graded relevance on IR effectiveness. His graded relevance scale had four levels, namely highly relevant, fairly relevant, marginally relevant, and non-relevant. For instance, this graded relevance scale is employed in Study II to scrutinize the effects of fallibility at different relevance levels. In addition to the challenges mentioned above, relevance feedback depends on initial result ranking. This can be problematic because a user's information need or knowledge state changes, since users usually learn even from the snippets of the search results presented with every document surrogate. Moreover, users can only indicate relevant documents, if such documents appear on the result page. On the other hand, if the user's information need is already fulfilled through the first result page, the necessity of applying RF is dissolved (by itself). In Study I we analyzed when to apply RF based on the precision of first result page. Apart from these, users have to assess every document individually. If user's fuzzy information need, imperfect knowledge state, and difficulties ascribed by diverse user interface issues, e.g., misleading snippets, are taken into account, it is not a big surprise that user can err (Vakkari & Sormunen, 2004) during relevance feedback. Consequently, the human fallibility in providing RF is simulated in Study II to measure the effect of fallibility on RF effectiveness.

However, users are not always ready to supply feedback information to the IR system explicitly even if they have a positive attitude towards giving relevance feedback (White, 2011). In such cases, IR systems may exploit implicit relevance feedback, which can be beneficial for current web search engines (Joachims & Radlinski, 2007; White et al. 2002).

2.3.2 Implicit Relevance Feedback

Because explicit RF has time and effort implications for users of an IR system, they may be reluctant to provide RF explicitly. A remedial action lies in user behavior during interaction with the IR system. Implicit RF does not require explicit feedback action from users, so in lieu thereof user's interaction with the search system will be observed and the user’s information need will be prognosticated from user behavior. For example, user’s dwell time on search result page and especially on some particular documents, saving or printing any result documents for further reuse, selecting, referencing or commenting any document could bring the required evidence for RF (Kelly, 2005; White, 2011).

Implicit RF is classified according to the user’s intent by Oard and Kim (2001) into four categories: first, “examine” behaviors, where users read, listen, view or select a document; second, “retain” behaviors, where users bookmark, save, or print a document; third, “reference” behaviors, where users give a reference to a document or its parts by replying, linking, or citing; finally fourth, “annotate” behaviors, where users annotate the documents by marking up, rating, or liking. However, not all categories can be leveraged during an online search, because IR search systems have limited access to users’ actions on document pages. Such limited access to user’s actions on various pages can be gained by search systems via “like” buttons, advertisement banners or similar constructs to follow the user click behaviors. Moreover, implicit relevance can be combined with explicit relevance to increase the effectiveness of IR systems and to corroborate the understanding of users’ needs by IR systems.

Furthermore, user profiles, which are based on users’ search history and preferences, could also be created by the search system, and can be incorporated into the search result building process as a way of implicit RF.

Although explicit relevance in our experiments is assumed and simulated, relevance information could be harnessed by diverse user behaviors in practical operational environments. Nevertheless, certain aspects of user behaviors are brought into simulation settings. Those aspects of user behavior are discussed in more detail in Section 3.2.

2.3.3 Pseudo-Relevance Feedback

Even though the usefulness of explicit relevance feedback accomplished with the query expansion technique was reported by Ruthven and Lalmas (2003), user’s reluctance to provide feedback remains a thorn in one’s side. Consequently, implicit RF can alleviate this burden to some extent. Yet another approach to utilizing RF, pseudo-relevance feedback (PRF) (Ruthven & Lalmas, 2003), assumes the top-ranked documents produced by an initial query to be relevant. PRF also avoids the user's explicit feedback. Before the first result page is presented to the user, which in case of explicit feedback is compulsory, PRF can hopefully be applied to improve search results before presenting them to the user. However, if the top-ranked documents are non-relevant, they cause query drifting, and can decrease the effectiveness of the IR system instead of increasing it. Nevertheless, IR experiments with PRF have shown that it improves the system outcome slightly. Moreover, to counteract the problem of query-drift, Järvelin (2009) and Lam-Adesina et al. (2001) introduced the query-based summarization of top-ranked documents in PRF and demonstrated the advantages of their approach. In addition to that, PRF can be combined with RF, at least after the first page is presented to the user to obtain implicit and/or explicit relevance feedback.

In Study I, we simulated a combination of PRF and explicit RF, in other words we applied our novel approach to the RF process on top of PRF results, and showed how this classification method can improve search effectiveness.

2.4 Applying Classification Methods for Relevance Feedback

Instead of selecting a conventional path to apply RF by expanding the initial query with additional keywords extracted from the relevant and non-relevant documents, or by modifying query terms weight (Ruthven & Lalmas, 2003), we opted for using classification and clustering methods (Sebastiani, 2002) to distinguish the relevant documents from non-relevant ones in subsequent results after RF provided by the user on the initial results. In order to train a classification or clustering algorithm, we assumed that a simulated user indicates the relevant and non-relevant documents on the first result page. The first result page was constructed by applying PRF to the initial query results, using the Lemur/Indri PRF algorithm (Strohman et al., 2005; Lavrenko & Croft, 2001); before the first result page is presented to the simulated user. Then the simulated user supplies the simulation program with relevance information based on the recall base of the collection used. Having trained the algorithms with two sets of documents representing relevant or positive ones, and non-relevant or negative ones, and built the classification model, the important question is whether the search program is able to discern the documents as positive or negative ones further down in the result list. Thereafter, the negative documents (or non-relevant ones) from the result list are removed and the relevant ones are shifted forwards to vacant positions in the result list.

However, before even starting with the application of the classification or clustering the subsequent documents after the first page results, one should ask the question: When should be the RF applied? Are we able to improve the results with the help of classification? Is it necessary or possible to improve the second and third page results at all, when the first page results either already satisfy the user's information needs or have no relevant documents? In order to answer these questions, it is necessary to analyze the precision of the first and subsequent result pages.

The next section briefly summarizes the main ideas of classification and clustering methods which are used in the present thesis. Further, the term space reduction algorithms, which are utilized during the classification process, are described. At end of the section, Learning to Rank is compared to the proposed RF classification approach, because both approaches apply machine learning methods and are therefore related.

2.4.1 Classification and Clustering Methods Used in the Present Thesis

The following classification and clustering methods are applied in the present thesis: Naïve Bayes classification method, K nearest neighbor (KNN) algorithm, KMeans algorithm and Support vector machines (SVM) classifier (Joachims, 1999; Manning et al., 2008, pp. 253-376; Sebastiani, 2002). These methods are selected because they are the most common and state-of-the-art methods in research and practice.

The Naïve Bayes classification method is based on Bayesian theory. This method tries to express posterior probabilities with prior probability and likelihood. Thereby, the probability of a document belonging to either the positive set or the negative set is defined by the highest posterior probability, which any of the sets produces with respect to the document being classified.

Again, the naïve assumption is also made here and the independence of words from each other in a document is assumed. Even though the words in a topical text depend on each other, obviously this assumption does not harm the outcome of classification and/or retrieval process severely. The probability of a document belonging to a class can now be easily determined by multiplication of the individual probabilities with which each word occurs in the documents of each class. In addition, the prior probabilities can be estimated according to the proportion of documents in each class, namely positive and negative. Still, the probabilities of the words in pertinent classes should be estimated and they can be computed from the frequencies in the training sets. However, test documents can contain words unseen during the training phase, which results in zero probabilities. Therefore, smoothing functions are applied, which assumes small probabilities for every word. Laplace correction is one of the smoothing functions employed in the present thesis. In our experiments we preferred the multinomial model, where the frequency of the words in documents is taken into account. An alternative, the multivariate Bernoulli model, regards word frequencies as binary features, which means that the duplicate words are eliminated from document representations. The multivariate Bernoulli model produces competitive results where very short documents like tweets are classified (e.g., neglecting the word frequencies does not change much) (Manning et al., 2008, pp. 253-288).

The K nearest neighbor (KNN) algorithm first calculates distances between the test document and all training documents. Then KNN selects the *k* nearest neighbors, which are the *k* closest documents to the test document according to whatever distance metric is used. Then the class of test document will be decided by a majority vote of the selected documents. There are several distance metrics in the literature such as Euclidian distance, Minkowsky distance, Manhattan or City-block distance, and Canberra distance (Losee, 1998, pp. 43-75). These were also employed in our experiments. Finally, we measured the similarity of documents to one another with the Euclidian distance metric, because it achieved the best effectiveness in our experiments. Even though KNN is quite a successful classification algorithm, it requires comparison with every other document in the collection of negative and positive documents, which happens on the fly after the test document is submitted. In comparison to Naïve Bayes, KNN does not build a model beforehand.

The clustering algorithm KMeans is also exploited for distinguishing between the relevant and non-relevant documents in our experiments. The documents on subsequent result pages are clustered with the help of the KMeans algorithm, which starts with the centroids of the clusters of the relevant and non-relevant documents pointed out by simulated user and tries to assign the documents to pertinent clusters by selecting the nearest cluster centroid. Similarly to KNN, nearest cluster centroids are determined by comparing the distances between the document and the centroid. After every document is assigned to any one of the either clusters, cluster centroids are recalculated with those documents in each cluster. This process is repeated until either the cluster centroid positions do not change anymore or the preset maximum number of iterations is attained.

A support vector machine (SVM) is the state-of-the-art classification algorithm, which separates data points, or the result page documents in our study, by means of a hyper-plane in a multidimensional space. First, SVM tries to find a hyper-plane which maximizes the distances to the nearest training data points, or rather documents, of two classes. These nearest data points to the hyper-plane are named support vectors, which play a major role in developing the training model; all the rest of the training documents will be discarded after determining the hyper-plane. After building the training model, or defining the hyper-plane, test documents can be readily projected to either side of the hyper-plane, which determines the appropriate class (Joachims, 1999; Manning et al., 2008, pp. 319-348).

2.4.2 Term Space Reduction Algorithms

The purpose of applying reduction is either performance improvement or the reduction of noise introduced by high dimensionality. Reducing the number of features also reduces the number of calculations, which would otherwise be computed for the disregarded features. This in turn contributes to the performance of the classification algorithm. On the other hand, one should also notice that term space reduction methods consume processing power. The benefit of term space reduction may be the avoidance of noise introduced by high dimensional feature space. However, some of the classification methods like SVM can cope with the high dimensions very efficiently by regarding only the decisive features.

In order to reduce the number of dimensions in our experiments, we experimented with the following methods: mutual information gain, Kendall-Tau rank correlation coefficient, Pearson's chi-squared test, odds ratio, Spearman rank correlation coefficient, and Fisher's exact test (Banerjee & Pedersen, 2003; Sebastiani, 2002).

The term space reduction methods employed in the present thesis are defined in the appendix.

2.4.3 Learning to Rank vs. Relevance Feedback Classification Approach

Web search phenomena triggered the ranking studies in order for the search engines to better serve search results for web search engine users (Li, 2011). However, ranking is not only confined to document ranking in web search, but is also applied in collaborative filtering like product recommendation, machine translation, and meta-search, which aggregates search results from several search systems. Nevertheless, we focus on the document ranking creation in this thesis. Initially unsupervised ranking models or rather ranking formulas like BM25, Language Model of IR and PageRank formula (Manning et al., 2008, pp. 219-252 & 461-482) and their combinations have been exploited in the evolution of search systems. These algorithms are unsupervised because they do not require any training phase with labeled data, even though some collection-dependent parameters should be gleaned first. These algorithms extract some features from queries and documents in the collection in order to calculate a score for the documents for a given query. There is a number of features, among others term frequency, BM25 scores, and edit distance for various parts of the documents like title, anchor text, URL, and body, as well as the number of incoming links and PageRank score of the document or page (Qin et al., 2010). Moreover, implicit RF based on click-through data or explicit RF features can also be collected in order to be employed in these algorithms.

However, the static way of determining the document or page score by these methods or their combinations can be improved by machine learning approaches. These machine learning approaches include classification and regression. In a broad sense, any machine learning algorithm applied to ranking can be called a Learning to Rank algorithm. On the contrary, a narrower definition of Learning to Rank is associated with the machine learning methods for constructing ranking models in ranking creation and ranking aggregation. The former creates rankings; the latter aggregates the rankings of different systems. Learning to Rank aims to create better rankings for search results. The most relevant documents will be placed on the top positions in the result list. Learning to Rank methods first create a ranking model out of training data, which is a collection of queries and respective documents labeled as relevant. Then the most relevant documents for future queries are retrieved by engaging that particular ranking model (Li, 2011).

The studies on Learning to Rank have produced plenty of different methods. In the main, there are three major approaches to learning a ranking model: *pointwise*, *pairwise* and *listwise* approaches. The *pointwise* approach considers the request and the respective documents individually. Naturally, both requests and documents are represented by feature vectors, which are involved in building the ranking model. The group structure between the request and relevant documents is omitted. This approach transforms the ranking problem into a classification, regression, or ordinal classification problem. For example, a learned ranking model can produce scores, e.g., real numbers in case of regression, for documents with respect to a query; thereafter documents can be ranked according to scores. The *pairwise* approach converts the ranking problem to pairwise classification or regression. Likewise, the pairwise approach ignores the group structure between request and documents. A classifier decides the ranking order of document pairs. On the other hand, the *listwise* approach takes the group structure into account, in other words, the ranking lists as whole are utilized both in the training and prediction phase. A ranking model ranks the documents according to scores which are reckoned by the ranking model. The *listwise* approach requires new methods because the existing machine learning techniques cannot be directly employed (Li, 2011).

Besides these major approaches, there are query-dependent and multiple nested ranking approaches. Furthermore, there are many diverse implementations of

Learning to Rank methods, some of which are also employed by commercial search engine companies. For example, the pairwise approach LambdaMART performed best in the Yahoo Learning to Rank Challenge (Chapelle & Chang, 2011).

In a broad sense, our classification approach for applying relevance feedback may be seen as a Learning to Rank method, because we employ a machine learning technique for re-ranking the search results and we have the same goal as Learning to Rank methods. However, Learning to Rank methods principally try to build a single model from training data, which will be exploited for future similar queries. In contrast, our approach builds a classification model based on either implicit or explicit relevance feedback after user's individual query and examining the very first result page and applies this particular model only for the rest of the results, which have already been collected for this specific query.

3. Simulation of Interactive Information Retrieval

Simulations are based on models. A model represents the phenomena that will be replicated in simulations and further captures the essential components and interactions. In this chapter, we first give an introduction to modeling and simulation, and then describe modeling behavior factors in simulations. Finally, we discuss search environments and cost aspects in session simulation.

3.1 Introduction to Modeling and Simulation

Before IIR simulation, which is the main focus of this thesis, is described, modeling and simulation are introduced generally in this section. Simulations add one more knowledge building tool in addition to the theoretical and experimental tools. Simulations can be exploited to gain insight, validate models and experiments, predict the potential outcomes of system changes, test and evaluate systems, among others (Sokolowski & Banks, 2011, pp. 25-43). In other words, simulations are executed to conduct *what-if* analyses. These *what-if* analyses in turn can contribute to gaining insight on the one hand and solving problems on the other. Problem-solving simulations incorporate less uncertainty, whereas gaining insight simulations are naturally plagued with more uncertainty, because the models used in these kinds of simulations are neither complete nor even accurate with respect to reality, which is simulated and investigated. The more insight is gained, the more accurate the models, and consequently the simulations, become. As a consequence the gaining insight simulation models are ephemeral by nature. However, problem-solving simulations can serve for longer periods, because these simulations are usually parameterized and have a stable model. The following type of questions, which are given in the simulation book (Sokolowski & Banks, 2011, pp. 25-43), can be answered by problem-solving simulations: *What would happen if…?, How will* *a…?, Why would a…?, Can a…?, Does the…?, Should we…?* On the other hand, the gaining insight type of simulations can answer the following questions: *What has the greatest influence? How will X and Y interact? Is there a way to make X happen? Why has unexpected behavior X occurred? What new behaviors might emerge?*

By nature, the simulations in the present thesis are of the gaining insight type, and those general questions can be specified such as: What has the greatest influence on cumulated gain in a session? What is the influence of user interface on cumulated gain in a session? How will gain and time in an information retrieval session interact? Is there a way to achieve the best gain? Why do the traditional evaluation metrics deliver unexpected results when time is taken in to account? What kind of new behaviors for search process might emerge?

First, a simulation model represents a real event, phenomenon or system, which will be simulated and analyzed, and is expressed in a formal way, usually mathematically or as a computer program. Therefore, a model should approximate the real event or system closely, and reflect the important features of the real world from the pertinent aspects of the simulation. However, incorporating every feature of the real world into a model not only increases model complexity but also makes the simulations infeasible. Consequently, some of the salient features are selected for model building and the rest are ignored. The balance between realism and the simplicity of the model is one of the critical decisions the model builder has to face. Too much simplicity diverts the simulation from reality, which may cause drawing the wrong conclusions. On the other hand, too much realism may cause computational difficulties, e.g., either excessive requirements for memory and/or processing time, which in turn makes simulations prohibitively expensive.

Models can be classified on the one hand as *static* or *dynamic* with respect to time, and on the other hand as *deterministic* or *stochastic* with respect to input or output variables. While dynamic models take time into account, static ones do not. Similarly, the stochastic ones model the probabilistic values for input or output variables, whereas the deterministic ones regard those variables as fixed. In Studies II, III and IV, we utilized the following model types, namely static, dynamic, deterministic, and stochastic models (Maria, 1997).

Before a simulation based on a model is further run, it should be verified and validated. While verification ensures that the model complies with its specification, validation enforces the validity of the model, which means that the model imitates the real event or system genuinely (Altiok & Melamed, 2007, pp. 1-10).

After a part of reality is modeled and formally expressed as a model, the model can be executed or, in other words, simulated in an environment, usually on a computer. Simulation allows input variables of the model to be changed and the execution of the model to be repeated, and then the outcome of the simulation experiment to be analyzed. In this way, simulations can shorten or extend the real time of a real event into a simulated time, which can be much shorter or longer. Hence, simulations empower us to test and analyze hypotheses about a real system in a timely manner, which can save enormous costs and efforts in comparison to real setup. In some situations real events cannot even be repeated easily or are almost impossible to repeat, for example because of side effects, which in turn affect the outcome of the real experiment. For instance, the information retrieval system users learn during searching in IR experiments, therefore the same task cannot be assigned to the same user for analysis of variations of the various task variables.

The simulations can be realized either as stand-alone programs, which run independently, or as integrated simulations, which are embedded into the real system. The stand-alone simulations can be classified according to application areas, for example: Training, decision support, understanding, education & learning, and entertainment. Further, the simulations can be classified according the user point of view, namely users or the researcher. While plain simulation users are more interested in problem-solving issues which they encounter day by day, for example during training, decision-making or entertainment, researchers rather try to gain insight into peculiar problems (Sokolowski & Banks, 2011, pp. 25-43). The simulations performed and analyzed in the present thesis can be regarded as stand-alone and simulations for understanding, because they facilitate hypothesis testing about the user behavior in IIR.

The following two sections address human factors in simulations, especially IIR simulations, and IIR session simulation. The first section models the fallible user during RF, further defines query modification strategies, and describes the scanning and assessment behavior, which are employed in the experiments in the present thesis. Finally, the scanning strategy of search results and the frustration of the user during search result scrutinizing are delineated. The second section specifies the simulation environments, and describes the cost factors in simulations.

3.2 Modeling Behavioral Factors in Simulation

One of the main focuses of this thesis is the modeling of human behavioral factors for IIR simulations (Azzopardi et al., 2011; Clarke et al., 2013). In order to achieve a realistic behavior representation, the observation of human subjects during information retrieval interaction is indispensable. An information need causes searchers to initiate a search process, which often consists of multiple queries in a session depending on the type of the search and the availability of documents. While traditional IR typically assumes a long query with persistent scanning of a long list of search results, we simulate more complex sessions based on user interaction with an IR system. Thereby, sessions may consist of multiple queries and/or user feedback. For the former, the searcher poses multiple queries one after another, and for the latter the searcher may give some feedback to the IR system, which utilizes the feedback to improve the search results before presenting them as an enhanced result list to the searcher. As one might expect, the searcher typically scans the search results and assesses the quality of the snippets before clicking the document links. After inspecting the respective document, the searcher then assesses the relevance of the document and judges whether the document fulfills their information need at least partly. This complicated process can be dissected into subtasks like scanning a snippet, clicking a document link, assessing the relevance of a document, and reformulating a query. These subtasks represent certain actions users apply during a search process. Each subtask is associated with certain effort that users expend. This effort may have many aspects and depends on the strategic decisions of the searchers. One way of measuring the effort can be realized in terms of time, which users spend performing the particular subtasks (Azzopardi, 2011).

However, searchers are human, and as we know, humans are fallible. We continuously make mistakes in every phase of the search process. Starting with misunderstanding the (work) task, searchers may type in misspelled search keys, or judge the relevance of the snippet or document incorrectly and consequently may give ineffective feedback. In spite of human fallibility in IIR, there is little research that takes the consequences of fallibility into account.

In this section, we first define the fallible user models for RF simulations to evaluate the effects of fallibility in RF experiments, whereby fallible humans are simulated according to some probability distributions. Second, we investigate the effectiveness of query modification strategies observed in real life. Then, we analyze the scanning and assessment behavior and their characteristics in terms of deterministic and stochastic ways. Finally, humans have limited scanning endurance, especially when the search results are of low quality; they either reformulate their query or give up the search session altogether. In order to model this human aspect, we discuss a novel frustration formula in the last subsection, which models the user's dedication to a search session.

3.2.1 Fallible User Modeling for Relevance Feedback Simulation

Modeling users for RF simulation requires several considerations regarding users' readiness to browse the initial search results and to give feedback, the level of relevance of the RF documents and users' fallibility during relevance judgments about documents. The first three points are addressed by Keskustalo and colleagues (Keskustalo, 2010; Keskustalo et al., 2006) by defining a user model. However, the last point, fallibility during relevance judgments, is a novel approach to user modeling in RF simulation. The motivation for this point comes from the literature, where for example Turpin and colleagues (2009) and Vakkari and Sormunen (2004) discovered erroneous relevance judgments of searchers.

Because we simulate the relevance judgments of users without resorting to their real judgments, we look for an alternative source of relevance judgments. Since the recall base of the test collection indicates the topical relevance level of the documents for respective query, we exploit the recall base as the source of relevance judgments for RF simulation. The simulation is conducted in the following way: Initial query results are scanned; each document on the ranked result list is checked against the recall base of the respective topic to obtain the relevance judgments.

When the recall base is applied as it is, i.e., relevance judgments are obtained directly from the recall base, this represents the deterministic case. We applied the deterministic case in Study I, where we accepted the topical relevance judgments about documents as given by the recall base. Thereafter, classifiers are trained to separate the relevant and non-relevant documents. However, does accepting relevance judgments of the recall base as gold standard reflect the real behavior of information system users? First of all, the recall bases which are utilized in experiments are generated by IR experts or by various persons who may have a dedicated task to develop test collections (Voorhees & Harman, 2000). On the contrary, we assume a typical information searcher. Actually, even assuming that users will agree with the expert's opinion about document relevance would be naïve, yet we also accept this assumption in our experiments; otherwise it would not have been possible to conduct the experiments and compare them with past ones.

Despite the fact that users usually judge correctly the relevance of a document, they can make mistakes between adjoining relevance levels (Vakkari & Sormunen, 2004). A probability distribution of making mistakes during feedback can be constructed. For example, according to such a distribution, a user may assess a document that is assessed by experts as fairly relevant, e.g., with 10% probability as non-relevant, 20% probability as marginal, 50% probability as fair (correct) and 20% as highly relevant.

We defined fallibility scenarios, which are employed in Study II for RF to construct models for fallible users. Our fallibility scenarios range from 100% correct judgments to completely random judgments. Although the probability distribution values are arranged with more probability mass to the correct relevance level and its neighboring levels, only some of them obey the normal distribution according to Shapiro-Wilk test[2], because the rest are limited by the far edge of relevance levels.

In addition to the systematically varied distributions, we designed one based on empirical observations by Vakkari and Sormunen (Vakkari & Sormunen 2004; Sormunen, 2002). They discovered that searchers are capable of recognizing highly relevant documents quite correctly but tend to err when dealing with the marginal and non-relevant documents. Therefore, in this empirically grounded distribution,

[2] "R: A language and environment for statistical computing" retrieved May 8, 2014, from http://www.r-project.org/

presented in Table 1 and labeled “0.50-0.80”, the probabilities are more peaked – 80% correct – for fairly and highly relevant documents, and flatter – only 50% correct – for non-relevant and marginally relevant documents.

Table 1. Fallibility probability distributions

Fallibility		Human Judgment Probabilities			
Scenario		**n**	**m**	**f**	**h**
0.50-0.80	**n**	**0.5**	0.4	0.1	0.0
	m	0.4	**0.5**	0.1	0.0
	f	0.0	0.1	**0.8**	0.1
	h	0.0	0.0	0.2	**0.8**

In more detail, the sample fallibility scenario labeled “0.50-0.80” in Table 1 defines respective assessment probabilities across the relevance levels. The row and column headers represent the relevance levels of non-relevant (n), marginal (m), fair (f), and highly relevant (h) documents. The row labels n to h represent the true document relevance labels as given in the test collection. The column labels n to h represent the (simulated) fallible human relevance judgments. The human judgment probabilities are given in the respective cells. The gold standard for RF would always deliver correct judgments on document relevance during the feedback process – that is, probability 1.0 along the diagonal in the table and other probabilities equal to zero. In the empirically grounded scenario of Table 1, the judgment probabilities approach the correct judgment for fair and highly relevant documents but deviate more from correct judgments for non-relevant and marginally relevant documents.

Further, we defined three more fallibility scenarios (see Study II, not shown in Table 1) which are motivated by the exploration of the effects of progressively increasing fallibility. Accordingly, we decreased the probability values of the sets systematically from fairly consistent judgments towards entirely random ones. Obviously, simulated relevance judgments are affected by random decisions made according to probabilities. As a consequence, RF effectiveness likely fluctuates in relation to random decisions. Therefore, we run each RF experiment multiple times in accordance with the Monte Carlo simulation approach (Altiok & Melamed, 2007, pp. 11-22) and report the average effectiveness.

3.2.2 Query Modification Strategies

Interactive search sessions can be characterized simply by querying and scanning iterations. While reformulating a query, various users prefer diverse strategies (Fidel, 1985). We examined some of these strategies, with which users achieve their goals under the constraint of the overall available session time. The procedure to reformulate a query can be defined in terms of query modification (QM) strategy.

First, we naïvely assume that a list of individual words {w1, w2, w3, w4, w5} is available for each particular topic, even though real life searchers learn from snippets seen and documents inspected (Vakkari, 2000). This can be seen as part of the simplification of the model for the simulation purposes. Nevertheless, to increase the realism of the experiments, we let two groups of test persons, students and researchers, suggest the search keywords for a set of topics. QM strategies determine how elements from this list are selected to form either an initial query or subsequent queries. In other words, the QM strategy defines how to form a sequence of queries (Keskustalo et al., 2009).

In Study III, we generated all possible query sequences with the permutations of the available individual search keys, and scrutinized their effectiveness. One should also note that the number of possible QM strategies becomes very large even with five search keywords and a limited number of queries per session. Thus, we limited the number of queries to three, which reflects real life search behavior (Jansen et al., 2000; Kamvar & Baluja, 2007; Yi et al., 2008). Besides, the required computation time for large number of queries per session would not be viable because of our limited computing resources. Still, it would be quite interesting to look at the lengthy sessions and their characteristics.

On the other hand, we paid special attention to five idealized versions of QM strategies, which have been employed by real users and reported in the literature (Keskustalo et al., 2009). We call them prototypical QM strategies. They are based on term-level changes. Consequently, we can only observe a limited number of queries per session by prototypical QM strategies, which also reflect real life behavior. According to a study by Jansen et al. (2000), the typical length of a search session is about three queries, and users employ 2.21 keywords per query on average. The prototypical strategies[3] are:

S1: an initial one-word query (w_1) is followed by queries which replace the word with the next one in the available list.

Q_1: w_1 -> Q_2: w_2 ->Q_3: w_3 ->Q_4: w_4 ->Q_5: w_5

S2: an initial two-word query (w_1 w_2) is followed by queries which replace the second word in the initial query with the next one from the available list.

Q_1: w_1 w_2 -> Q_2: w_1 w_3 -> Q_3: w_1 w_4 -> Q_4: w_1 w_5

S3: an initial three-word query (w_1 w_2 w_3) is followed by queries which replace the third word in the initial query with the next one from the available list.

Q_1: w_1 w_2 w_3 -> Q_2: w_1 w_2 w_4 -> Q_3: w_1 w_2 w_5

S4: an initial one-word query (w_1) is followed by queries which extend the previous query with the next search word from the available list.

Q_1: w_1 -> Q_2: w_1 w_2 -> Q_3:w_1 w_2 w_3 -> Q_4:w_1 w_2 w_3 w_4 -> …

S5: an initial two-word query (w_1 w_2) is followed by queries which extend the previous query with the next search word from the available list.

Q_1: w_1 w_2 -> Q_2: w_1 w_2 w_3 -> Q_3: w_1 w_2 w_3 w_4 -> …

3.2.3 Scanning and Assessment Behavior

After posing a query to a search system, a user may scan one or more documents before formulating the next query or ending the search session. If the search process is simply split into scanning and querying, after a single query Q_i a sequence of one or more document snippets may be scanned (s_{ij}: scanning the j^{th} document for query Q_i):

Q_1->s_{11}->s_{12}->s_{13}->…

[3] In Study IV we omitted strategy S4, therefore strategy S4 in the paper represents strategy S5 in this summary.

A user can scan a varying number of document snippets after posing any particular query to a search system during a search session. This results in a vast number of possible querying-scanning sessions, e.g.:

Q_1->s_{11}->Q_2->s_{21}->Q_3->s_{31}-> … or

Q_1->s_{11}->s_{12}->Q_2->s_{21}->… or

Q_1->s_{11}->s_{12}->s_{13}->Q_2->s_{21}-> s_{22}->Q_3->s_{31}->… etc.

A typical search session continues until the user's information need is at least partially satisfied and/or all the time allocated for the session is consumed or the user has no further ideas for a new query or is unwilling to produce new queries. The scanning lengths may fluctuate for many reasons depending on the user's belief in the success of the current query (Carterette et al., 2011) and the user's accumulated total gain during the whole session. Therefore, we analyzed the properties of optimal and suboptimal interactive search sessions for given time constraints. For the analysis, all possible sessions, which were formed by all combinations of scanning lengths using a sequence of available queries for each topic, were produced. In simulations we confined our focus on the first result page, assuming ten documents on a page, because only a few top documents are often inspected by users in real life (Jansen et al., 2000; Ruthven, 2008). Therefore, the top ten document snippets were taken into account when the scanning length combinations were built. In Study III, we utilized all five QM strategies (see Section 3.2.2) and varying scanning lengths in order to find the best-performing combination.

Furthermore, we can elaborate on the search process with more subtasks such as scanning the snippet, clicking the link, reading the linked document, and judging its relevance. Every subtask may be associated with a cost, e.g., in terms of time. Consider the handling of a single query Q_i again.

Q_1->s_{11}->c_{11}->r_{11}->j_{11}->s_{12}->s_{13}->c_{13}->r_{13}->j_{13}->…

Here s_{ij} stands for scanning j^{th} snippet for i^{th} query, c_{ij} clicking on the snippet, r_{ij} reading the linked document, and j_{ij} judging its relevance. If the searcher clicks on every snippet on the search result page, and reads and judges every document which is clicked, the cost of this session manifests as the sum of action costs (e.g., assuming the first and third documents are relevant):

$qc_1 + sc_{11} + cc_{11} + rc_{11} + jc_{11} + sc_{12} + sc_{13} + cc_{13} + rc_{13} + jc_{13} +…$

Assuming a particular number *n* of keywords is either available or searchers are ready to produce, it is possible to generate 2^n-1 word combinations for distinct queries. Further, when a set of queries is available for each topic, the searcher can scan a varying number of document snippets after any query. Altogether, this leads to a great many possible querying-scanning-reading-judging sessions.

However, document snippets are not always informative and searchers may overlook them (Ruthven, 2008; Turpin et al., 2009). This may cause the searcher to skip some of the snippets and documents, which should be read and assessed otherwise. Furthermore, the relevance judgments of the searchers may be different from experts' opinions or topic relevance assessors. In order to simulate this behavior we selected the probabilities given in Table 1. Table 2 shows the clicking and assessment probabilities by the relevance degree of the documents. For instance, the simulated searcher will click the snippet of a non-relevant document (of relevance degree 0) with the probability of 27%. Top-ranked but still non-relevant documents may mislead the searcher to click the link because of the apparent snippet relevance (Ruthven, 2008). Further, a searcher may judge a non-relevant document with the probability of 20% as relevant. The probabilities increase toward highly relevant documents which are judged as relevant with the probability of 97% because searchers are able to readily recognize highly relevant documents (Vakkari & Sormunen, 2004).

Table 2. Action probabilities by document relevance degree

Feature of Behavior	Relevance Degree			
	0	1	2	3
Clicking Probability	0.27	0.27	0.34	0.61
Judgment as Relevant Prob.	0.20	0.88	0.95	0.97

We model two types of session behavior in respect of interaction with the result list for Studies III and IV, namely deterministic and stochastic behavior cases. In the deterministic behavior case, we assume that a searcher always decides to execute the subtask and there is no probability to skip the subtask. For example, a searcher will always scan all ten documents in the result list confined by the scan length constraint and available time for searching, will click all relevant document snippets in the scanned result list, as well as will judge them correctly (see Figure 3).

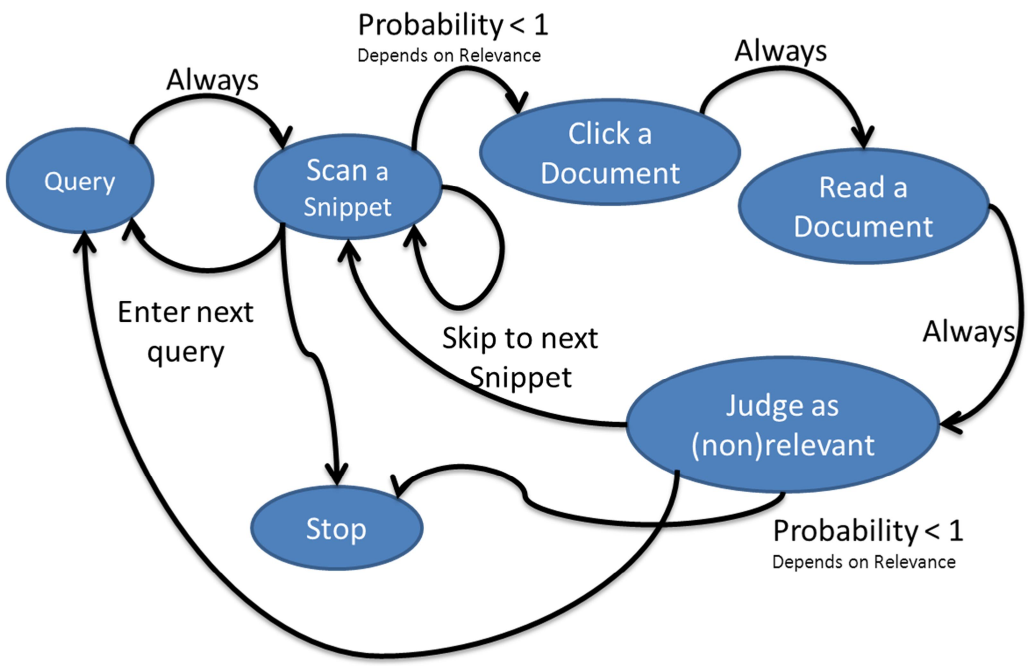

Figure 3. The simulation automaton depicts search session with subtasks

However, in the stochastic case, we assume more realistically that a searcher may err and sometimes may make the wrong decisions with some probabilities, as in real life. The simulation of stochastic behavior is established around scanning and assessment probabilities given in Table 2. For example, after posing a query to a search system, the searcher scans the result list and clicks some of those document links according to the selected probability values, as well as reads and judges them with some other probability (see Figure 3). We simulated sessions representing stochastic behavior with the help of the Monte Carlo approach (Altiok & Melamed, 2007, pp. 11-22). This means that we ran the experiments multiple times with random decisions according to the given probabilities, then averaged the outcomes of the experiments in order to achieve a more stable and robust analysis of the underlying phenomena.

3.2.4 Modeling Frustration

In real life, a searcher can scan one or more documents or links before frustration strikes, as a result of futile results, and then stops scanning further and reformulates the next query. Therefore, several factors affect the decision to continue scanning the results of a current query in a session: the search gain goal, the gain accumulated by the preceding queries in a session, the gain accumulated by a current query before the current scanning position, and the length of the current scan from rank one.

However, earlier studies (e.g., Carterette et. al. 2011), which model result scanning processes, assume single query sessions. This means that users are assumed to pose only one query and the proposed models predict at which rank users stop scanning the search results and quit the search process. In other words, they focus on the utility gained in a single query session. On the other hand, other researchers (e.g., Kanoulas et. al. 2011) modeled the multi-query sessions but did not consider either the search gain goals or the gain accumulated by previous queries in a search session. In summary, none of prior studies into scanning length modeling take multiple query sessions, varying gain goals, and simultaneously user's efforts as well as frustration into account.

Consequently, in the present thesis we modeled the user's frustration and contributed a novel formula for the scanning process regarding multiple query sessions in Study IV.

3.3 Session Simulation

3.3.1 Search Environments

First of all, Studies I-IV each had their own simulation environment, which emphasizes the particular aspects of that study and is restricted by the respective study objectives. In Study I we simulated relevance feedback on top results, and learned classifiers to classify the remaining results in order to improve IR effectiveness. Here, we resorted to the recall base of the collection (Voorhees & Harman, 2000) as the source employed to assess the topical relevance of documents on the first result page. Of course, the evaluation procedure makes use of the recall base too. Further, in Study II we again utilized a recall base, but we also introduced some noise into the user's interpretation of the recall base to represent the real life user's struggle concerning document and snippet relevance judgment. In Study III we created a simulation environment in which we examine the effects of various QM strategies in two interface scenarios. Below we describe this environment more precisely. Finally, in Study IV we expand our simulation environment to conduct experiments to study the behavioral factors in the search process. In the last two studies, we produced all the possible variations of some independent variables, such as scanning the search results under several constraints to obtain statements about the effects of the independent variables on the dependent one such as effectiveness, while certain variables like search environment were held constant. These variables will be discussed in more detail later.

For the session simulation in Study III, we first formally generate all possible sessions under constraints. We represented sessions as sequences of actions with costs, because the core of this study was about time aspects of different subtasks. For example, the tuple $<(a_1,c_1), (a_2,c_2), \ldots, (a_n,c_n)>$ is a session of n actions and each pair (a_i,c_i) in the session represents an action a_i and its cost c_i in seconds. The elementary action types and costs are:

- initial query, represented as ('iq', ic)
- query reformulation ('q', qc)
- document snippet scan ('s', sc)
- next page request ('n', nc)

The constraints are:

- MaxSLen, maximum session length in terms of elementary actions
- MaxSCost, maximum session cost (seconds)
- a session always begins with an initial query
- all queries (initial and reformulation) are followed by at least one snippet scan

Consequently, the shortest possible session can be formed by an initial action IA $= <(iq, ic),(s, sc)>$, consisting of an initial query followed by the scan of one snippet (with costs). To generate longer sessions, the possible subsequent elementary actions with costs are defined as the set:

$$NA = \{<(q, qc), (s, sc)>, <(s, sc)>, <(n, nc), (s, sc)>\}$$

Note here that the next actions are tuples of one or two elementary actions; a scan may appear individually, while a reformulation/next page requires a scan to follow.

Sessions are generated by concatenating the actions subsequent to the initial action. This operation generalizes over a set of session tuples S_i, denoted as:

$$\times_{i=1\ldots n} S_i = <<\ldots <<S_1, S_2>, S_3>, \ldots>, S_n>.$$

The cost of a session S can be determined, informally, by the sum of its action costs. More formally, we derive this cost by the function s-cost as follows:

$$s\text{-}cost(S)= \Sigma_{(a,c) \in S}\, c$$

Notably, the definition of the set membership operator was enhanced from sets to tuple components in an obvious way. For example, the cost of the session S1 = <('iq', ic),('s', sc), ('q', qc),('s', sc)> is *s-cost*(S1) = ic+sc+qc+sc.

To generate sessions, we first generate all sessions up to the maximum number of actions MaxSLen. This session set is MLS:

$$MLS = \cup_{i=1\ldots MaxSLen}\{<IA, \times_{j=1\ldots i}\, ac_j > \mid ac_j \in NA\}$$

We then select the subset of sessions fulfilling the time constraint MaxSCost and the scan length constraint. Note that this approach does not define the query contents or modifications in sessions (see Section 3.2.2). However, it keeps them within constraints and guarantees that the last action is a document snippet scan.

For session simulations in Study IV, we first generated all possible sessions under constraints as in the Study III. However, this time we refined the behavior aspect during search process further. Namely, we introduced more action types such as “click a link”, “read a document” and “judge document relevance”. Moreover, we integrated all query strategies into sessions, while paying special attention to prototypical QM strategies (Keskustalo et al., 2009). Thereby, all possible queries were constructed by a combination of keywords suggested by the real users for each topic (Keskustalo et al., 2009). Furthermore, sessions were executed many times in the stochastic case to encounter the randomness of the decisions concerning the actions. Again, we represented sessions as sequences of actions with costs because of time constraints. Nevertheless, the main focus of these simulations was the effectiveness of prototypical QM strategies employed by real users compared to all possible QM strategies under constraints. In the deterministic case, we ran more than a million sessions for each experiment, while for the stochastic case we ran more than a billion sessions for each experiment.

The next subsection explains and justifies the cost aspects used in the studies.

3.3.2 Cost Aspects

The effort to formulate a query, to scan the result list, to read documents and to judge the documents can be characterized by cost, or rather in terms of time expended (Azzopardi, 2011). Average subtask costs, which are utilized in our experiments, are given in Table 3. Thereby the scenario, which depends on the search environment such as the access device, determines the absolute cost. Empirical studies show that it takes significantly longer to enter queries using a small smartphone keypad than it does using an ordinary keyboard (Kamvar & Baluja, 2007). Two scenarios, i.e., a desktop PC scenario (PC) and a smartphone scenario (SP), are designed to study the effects of subtask costs under overall session cost constraints. These scenarios have different subtask costs, because the properties of the devices partially determine the user's effort to accomplish the subtasks (Kamvar et al., 2009; Smucker, 2009).

Table 3. Average subtask costs used in Study IV (in seconds)

Session subtask	**Costs**
Entering a query word	3.0
Scanning one document snippet	4.5
Reading and evaluating one document	30.0
Entering the relevance judgment	1.0

Obviously, forming queries under different QM strategies S1 – S5 (see Section 3.2.2) also leads to very different costs. All queries in strategies S1, S2, and S3 have a fixed query length in sessions (one, two or three words, correspondingly) while in strategies S4 and S5 the queries grow longer. In real life the typing speed is affected by, e.g., the experience and knowledge of the person, the size of the keyboard, the layout of the keyboard (e.g., nine-key multi-tap vs. QWERTY keyboard) (Kamvar & Baluja, 2007; Karat et al., 1999) and whether a predictive text feed is available and used. We derived the cost values in PC and SP scenarios regarding the initial query cost and the subsequent query cost from literature (see Table 1 in Study III) (Kamvar & Baluja, 2007).

The query costs in S1 – S5 in the desktop PC case are based on the typing costs of 3.0 seconds per word. The corresponding smartphone case costs are based on the article by Kamvar and Baluja (2007). The authors performed a large-scale log analysis of mobile phone usage and observed that an average smartphone query length was 2.56 words and the average query-entry time was 39.8 seconds (average typing cost of 15.5 seconds per word). We assumed in our simulations that the cost of adding one word to a query (that is, extending one-word query and extending two-word query strategies, S4 and S5) or replacing one word at the end of the previous query (that is, one, two, or three-word query strategies) is a constant depending on the scenario.

The information processing of humans can be approximately described by perceptual, motor, and cognitive systems (Card et al., 1983, pp. 23-100). The document snippet scanning costs in real life are affected by the costs accumulated by the above-mentioned systems. However, in Studies III and IV we assumed that the document snippet scanning cost is constant in both scenarios and across the QM strategies. Moreover, we excluded the eventual thinking time in producing query words, which can be interpreted as a modeling artifact because of simplification of the real world.

4. Evaluation of Interactive Information Retrieval

Without measuring the performance of systems and the outcomes of experiments, real progress in scientific pursuit cannot be achieved. Therefore, this chapter briefly introduces the evaluation methods (Catarci & Kimani, 2013), which are applied in our experiments. Further, the statistical methods, which are utilized to show the statistical significance of the experiment results, are shortly described.

4.1 Rank-Based Evaluation

Information retrieval has a long tradition in measuring IR system performance with respect to ranking of the search results. From the system-oriented view of IR, the most important aspect is the rank of documents which are returned as a query result. Precision, the proportion of retrieved relevant documents to the retrieved documents, and recall, the proportion of the retrieved relevant documents to all known relevant documents, are the very first measures which are applicable to the results of IR experiments. For instance, mean average precision (MAP) is a widely-used measure to compare systems with each other. MAP is calculated by averaging the precision values at the ranks of retrieved relevant documents of a query result, and thereafter the mean value of all query averages. Another recently popular measurement, cumulated gain (CG) (Järvelin & Kekäläinen, 2000), is based on the gain that every document contributes. The gain of a retrieved document is usually associated with the relevance level of the document. Further, ranks of the retrieved documents are taken into account by discounting the gain factor according to the position of the documents in the result list. This results in discounted cumulated gain (DCG) (Järvelin & Kekäläinen, 2002). However, in order to accomplish the comparability between different systems or experiments, the cumulated gains should be normalized; indeed the normalized discounted cumulated gain (NDCG) introduced by Järvelin and Kekäläinen. (2002) has become one of the most widely-applied evaluation measures in IR domain. DCG is normalized with the help of ideal discounted cumulated gain, which can be calculated by summing the discounted gains of known relevant documents for each query in descending order up to the rank where NDCG value is required. Across the queries of an IR experiment, NDCG values are averaged by taking an arithmetic mean in order to obtain a final NDCG value.

4.2 Time-Based Evaluation

IR evaluation is traditionally considered a rank-based process. However, when the time a user expends during a search session is taken into account, traditional metrics are inadequate for evaluating the search results. Because traditional metrics are time agnostic, a user's effort in terms of time is entirely omitted in the evaluation process. However, there have also been research efforts which introduce the time dimension into the evaluation process. For example, Dunlop (1997) suggested "time-to-view" graphs, which incorporate user interface and system as well as the temporal issues into the same framework in order to evaluate search engine effectiveness. According to Dunlop, "time-to-view" graphs offer a single presentation, which enables researchers to compare the interface and effectiveness changes.

Another research effort to introduce the time factors into the traditional Cranfield setting was conducted by Smucker (2009). He tried to improve the traditional evaluation with the use of the GOMS model (Card et al., 1983, pp. 139-192), which stands for goals, operators, methods, and selections. Further, he proposes an IR user model, which incorporates the sequence of actions performed during the searching process, such as typing, clicking, evaluating a snippet summary, and waiting for the results to load. All these actions can be associated with times and probabilities, with which users perform the actions. For instance, whether a user will click on a relevant document surrogate is defined by a given probability. He studied simulations to show the impact of changes in the information retrieval interface on user performance, which was determined by the number of relevant documents read within a given time frame. Moreover, for IR evaluation he suggested a time-biased gain metric, which captures some aspects of user behavior by regarding the search process (Smucker & Clarke, 2012a; Smucker & Clarke, 2012b). The suggested metric is calibrated through a user study for stochastic simulations. Furthermore, in a subsequent article (Smucker & Clarke, 2012c), the authors simulated different types of users by modeling user variance in time-biased gain in order to estimate the expected number of relevant documents that a user will collect while examining a single ranked result list. Still, their experiments were limited to single query sessions.

Azzopardi (2011) approached interactive IR as an economical problem and examined the trade-off between querying and browsing while holding search utility constant, computed in terms of normalized CG, at a certain level. He employed a user cost function in order to determine the search strategy, which keeps the minimum cost at the constant utility level to a user. The suggested user cost function takes the cost of querying and browsing into account, and is proportional to the number of queries issued. The time expended for querying and browsing is utilized to define the relative cost. Azzopardi (2011) claims that the user cost function estimates the relative cognitive effort of querying and browsing and his approach offers a reasonably fair comparison between strategies.

In our experiments, we take the user's effort as a variable represented by time into account. Consequently, we propose a new time-based evaluation approach. Nevertheless, we utilized the cumulated gain (CG) over time to compare the effectiveness of sessions as well as search strategies because of some very interesting peculiarities with traditional metrics, such as MAP and NDCG (see Chapter 6).

4.3 Statistical Methods

Parametric and non-parametric statistical methods are applied in order to identify statistically significant differences between the proposed and the state-of-the-art techniques in IR experiments. Thereby, parametric methods assume some statistical distributions to judge the differences between the examined algorithms, which are tested for significance in preset level of confidence. On the contrary, the non-parametric methods do not depend on such statistical distributions; they rather apply the rank-based calculations for statistical tests. The following two statistical methods, namely the t-test and the Friedman test, are used in the present thesis to assess the significance of differences between the algorithms in our experiments. Therefore, the t-test and the Friedman test (Hill & Lewicki, 2007, pp. 15-40; Conover, 1999, pp. 367-372) will be briefly described below.

The t-Test, a parametric statistical test, analyzes two data samples and estimates whether the data samples are drawn from the same distribution. Student's t distribution is the underlying distribution for the t-test. In other words, the t-test examines the equality of the means of the two normally distributed samples with unknown variances. If the sample size is large enough, the normality assumption can be relaxed to some extent. Another requirement of the t-test is that the variances of the two data sets should not be too different.

However, when the data samples do not follow the normal distribution, the non-parametric alternatives are more proper than parametric ones. As a non-parametric alternative, the Friedman rank test is selected for the current study for the cases where the data do not follow normal distribution. The Friedman test executes two-way analyses of variance by ranks. Besides, the Friedman test can handle more than two data samples, which occurs in our studies. In brief, the Friedman test first checks whether a significant difference between data samples exists, then calculates pairwise comparison between the data samples, which are produced by the diverse methods which are under evaluation. In order to accomplish the test, the Friedman test determines whether the data samples originate from the same population or populations with the same median, by determining the probability of divergence of rank totals of the samples from the rank values obtained by chance. The Friedman test is explained in more detail in Conover (1999, pp. 367-372).

5. Summary of Contributed Studies

In this chapter we present the summaries of the four contributed studies. We briefly explicate motivation, problems, approach, and data for each respective study. Then we present the research questions in a succinct form. Finally, we describe the research results of each study.

Studies I and II handle RF simulations, while Studies III and IV simulate session behaviors without RF. While Studies I and II handle single query sessions, Studies III and IV utilize multiple query sessions.

5.1 Study I: Effectiveness of Search Result Classification based on Relevance Feedback

Relevance feedback has been one of the research areas of system-oriented IR for a long time. It has been studied by utilizing either test persons or simple simulations. RF has been conducted through query reformulations with the help of PRF and/or intellectual RF (Ruthven & Lalmas, 2003). In Study I we performed RF in a novel way through the classification of search results after users' initial intellectual feedback. We simulated users' initial intellectual RF for our experiments in a comprehensive collection, namely the TREC 1-2-3-7-8 ad-hoc collections with 250 topics (Voorhees & Harman, 2000). We tried several classifiers, which are explained in Section 2.4.1. We also studied the effects of diverse term space reduction techniques for the classification process. Experimental results were evaluated by user-oriented metrics, P@20, P@30, NDCG@20, and NDCG@30. The following research questions (RQ) were set for Study I:

RQ 1: Given RF on top results of pseudo-RF (PRF) query results, is it possible to learn effective classifiers for the following results? What is the effectiveness of various classification methods?

RQ 2: How does classification effectiveness in RQ 1 depend on term space reduction and classification methods?

RQ 3: When should RF and classification be employed regarding the availability of relevant results in the initial Top 10?

In Study I we propose a novel approach to applying RF. Our approach trains classifiers with the help of simulated user-RF on top of PRF results instead of reformulating the initial user query by expanding with keywords extracted from RF documents. These classifiers are then applied to identifying relevant documents among the subsequent search engine result documents, which have not yet been presented to the user as a result list.

For the first RQ, our results indicate that the proposed classification approach can be applied effectively on top of PRF results. In both cases of title only (T queries) and ‘title-and-description’ queries (T+D queries), the proposed classification approach improves both the initial query results and PRF results, while the improvements over PRF results are smaller than the ones over initial results. This suggests that even though state-of-the-art search engines have so much evidence from long initial queries, the classification approach can still improve the results by learning through top document RF. All in all, the effectiveness of both the short and the long queries can be improved with classification approach. Furthermore, all tested classification methods provide statistically significantly better results over PRF and initial query results.

For the second RQ, further results indicate that term space reduction is no more effective than using the full feature set in T+D queries but that it provides a marginal boost in the shorter T queries. Although the best results for short queries are achieved either by a classification method other than SVM with term space reduction or SVM (Joachims, 1999) with full feature set; the differences between classification methods were minor and statistically not significant. The best methods for long queries were KNN and SVM with a full feature set, while they had only an insignificant advantage over the other classification methods without reduction. However, one should also note that SVM with all features performs quite well, because term space reduction is an integral part of this method. Reduction in SVM is applied by selecting the support vectors; in this vein, term space is implicitly reduced.

For the third RQ, we found that the classification approach should be applied when there is at least one relevant and non-relevant document in the initial result list. Regarding the searcher behavior, if the result list contains only relevant documents, searcher‘s information need is probably satisfied on the first page. On the other hand, learning a classifier with only positive or only negative documents complicates building classification models for document space. Moreover, our analysis points out the high correlation of P@10 with P@11-20/30 which means that if first result page has many relevant documents, the subsequent pages will have also many relevant documents, and vice versa, if first result page has no relevant documents, the subsequent result pages will have hardly any relevant documents. Consequently, high correlation between first and subsequent result pages supports the finding on where the classification effort should be focused.

Our findings are based on user simulation. We modeled searcher interaction during RF and assumed feedback on the Top 10 PRF search results. Realistically, we simulated that users browse the first page. However, the assumption of RF for all documents in the first page may be questionable regarding the observation in the IR literature on searcher behavior (Ruthven, 2008).

Finally, our findings indicate that this novel approach of applying RF is significantly more effective than PRF with short and long queries. This paper inspired us towards more elaborate models of user interaction in IR. Namely, we applied the ideas about user fallibility in RF in the next paper.

5.2 Study II: Simulating Simple and Fallible Relevance Feedback

In the previous study, relevance feedback was performed under laboratory conditions using test collections and a simulated deterministic searcher. In order to improve the realism of the experiments we designed a unique experimental setup in Study II. First of all, instead of title-and-description derived queries, we introduced realistically short queries that were suggested by real persons (Keskustalo et al., 2009). Second, we simulated human fallibility by providing RF, i.e., partially incorrect judgments about the documents in the feedback process (see Section 3.2.1). Third, we performed a user simulation with several evaluation scenarios. Finally, we employed graded relevance assessments in the evaluation of retrieval results.

The research questions were:

RQ 1: How effective are PRF and RF when employed on the results of short initial queries and shallow browsing?

RQ 2: Does RF effectiveness seriously deteriorate when RF is of progressively lower quality?

RQ 3: How does RF effectiveness in RQ 2 depend on evaluation by liberal and fair vs. strict relevance criteria?

In order to study real world problems in a laboratory environment, we established a simulation environment, in which a simplified model of the real world is utilized to conduct the experiments. This motivates our present study in which we model user interaction features during RF and vary them systematically.

At first, the relevant features of real world searching were studied in order to fulfill the requirements for more realistic simulations. Day-to-day observations corroborate that interaction in real life IR is indispensable. Besides, individual users interact with information retrieval systems differently. However, a typical real life searcher interaction can be characterized as being simple and error prone, or more specifically, searchers try to achieve the best results with minimum effort, in other words with short queries as well as shallow browsing (for example at most the top 10 documents/snippets checked, rather than the top 1000) (Jansen & Spink, 2006; Jansen et al., 2000; Sakai, 2006). Because providing RF requires extra effort from searchers, they may be reluctant to give it (Ruthven & Lalmas, 2003). If they are ready to provide RF in order to achieve better results, they may make errors when judging the relevance of the feedback documents (Vakkari & Sormunen, 2004).

In our simulations for Study II, we employed (1) very short initial queries, namely one, two and three-word queries; (2) shallow browsing (assuming that at most the top 20 documents per query were inspected); and (3) we also defined the fallibility of the searcher during the providing of the RF. Fallibility was modeled according to several scenarios, assuming that searchers may err during the selection of feedback documents. These scenarios range from assuming perfect user judgments to completely random judgments. In addition, we define a scenario (see Section 3.2.1) based on empirical findings on the level of fallibility when the user attempts to recognize relevant documents belonging to various relevance levels (Foley & Smeaton, 2009; Vakkari & Sormunen, 2004). A total of five different fallibility scenarios were analyzed. All experiments were run multiple times in line with the Monte Carlo approach (Altiok & Melamed, 2007, pp. 11-22) with random decisions which obey the defined fallibility probabilities, and the results of all runs were averaged to infer reliable statements about the subject matter.

The evaluation of the experiments was based on user-oriented measures, P@10/P@20 and a traditional system-oriented measure, MAP. In retrospect, it would have been interesting to employ cumulated gain-based metrics and to compare the results accordingly. Since in real life users differ in their preferences considering satisfaction levels, we applied three distinct relevance levels. In other words, some users prefer finding even marginally relevant documents, while others want to obtain only highly relevant documents because their expertise in topics varies. Moreover, we decided to exclude the seen documents from RF results, which means we applied full freezing (Keskustalo et al., 2008), because users would not gain any benefits from seeing the same documents in the improved result list after expending effort to inspect them, regardless of their relevance level, in the first result set.

Regarding the first RQ, our results suggest that both PRF and direct user-RF applied by using query-biased summaries[4] are promising methods when very short initial queries are used. For the second RQ, as we expected that although increasing error level in providing RF progressively decreases the performance compared to perfect RF, it is still slightly better than the best-performing PRF. Surprisingly, RF with the empirical level of fallibility yields results that are close to perfect RF results. Considering the third RQ, assuming empirical fallibility and using user-oriented measures such as P@10 and P@20, RF performance systematically exceeds the performance of all short-query types (one, two and three-word queries) at a liberal level (i.e., even marginal documents are accepted as relevant). However, RF does not improve the performance of all short queries against PRF, when strict

[4] Query-biased summary process is depicted in Study II in Figure 1.

evaluation is required (i.e., only highly relevant documents are accepted as relevant). This may be part of the reason why RF does not prevail in real life.

Our findings suggest that completely random feedback is no different from pseudo-relevance feedback and is not effective in short initial queries. However, RF with empirically observed fallibility is as effective as correct RF and is able to improve the performance of short initial queries.

Next, we turned to focus on modeling the user characteristics during interaction with a search system. We also take user effort during interaction into account. We extended our experiments session dimension by undertaking multiple queries. In other words, we simulated direct query reformulation. Obviously, this strategy means that we do not study the RF process any more in the following studies.

5.3 Study III: Time Drives Interaction: Simulating Sessions in Diverse Searching Environments

In real life, users often conduct search activities by posing multiple queries during a search session (Jansen et al., 2009; White & Drucker, 2007), whereby searching consists of various cognitive, perceptual and motor subtasks (Smucker, 2009). During interaction with a search interface, users apply diverse strategies which affect their effort (cost), experience and session effectiveness. In Study III we suggest a pragmatic evaluation approach based on scenarios with explicit subtask costs. Furthermore, the effectiveness of diverse interactive strategies, namely query modification and scanning strategies, in two search environments, namely in desktop PC and smart phone search environments, was studied comprehensively. We simulated 20 million sessions in each environment to cover all possible interactive search scenarios that were possible within the study design. This in turn enabled us to analyze the effectiveness of the session strategies (see Section 3.2) and the properties of the best and worst performing ones in each environment.

We set the following three empirical and one methodological research questions (RQ):

RQ 1: How effective are the five QM strategies (S1 to S5, see Section 3.2.2) in terms of cumulated gain when we compare the Desktop PC and the Smart Phone (SP) scenarios under overall time constraint?

RQ 2: What are the characteristics of the best and the worst QM sessions?

RQ 3: How stable are the observed trends when the overall time constraint changes? Can we recommend QM strategies based on the PC and SP scenarios assuming a specific time constraint?

RQ 4: What is the proper evaluation methodology when time is part of the evaluation criteria?

In this study, we simulated various search scenarios on two different devices, a desktop PC and a smartphone, regarding diverse search subtask costs under an overall time constraint. Furthermore, the characteristics of the best and worst search sessions were explored. Because real life users have limited time to acquire the necessary information about their task and they use different devices for information access in different situations, our study has unquestionable user relevance, and consequently offers potential pragmatic value to the industry. Measuring the effectiveness of search systems from a user's point of view may reflect the user's interest more accurately, and thus increase the validity of the results achieved.

The first RQ was about the effectiveness of different QM strategies under time constraints. In the desktop PC scenario, when time is tight users cannot pose all possible queries, or utilize their entire search vocabulary. Instead, users may perform exhaustive scanning for a few queries posed. Short queries (strategy S1) perform worst in terms of session effectiveness, which is measured by cumulated gain metric. On the other hand, two or three-word queries clearly outperform the short queries. The same improvements in results can also be observed in strategies S4 and S5, when there is enough time to advance beyond the first query. When more time is available for searching, the initially weaker strategies catch up because users can scan more results, and the ranking of weaker strategies is not that critical. In the smartphone (SP) scenario, users have no time for long queries in a stringent time frame; therefore they must employ shorter queries and scan the weaker quality rankings. The effective strategies require a high query input cost; consequently they may not be applied at all. Again, the more time users have at their disposal, the smaller the gap between the effectiveness of best and worst sessions.

Regarding the second RQ, in both PC and SP scenarios and under stringent time constraints, the best sessions involved less queries and longer scans than the worst sessions. However, when more time is available, the differences between session characteristics in the PC scenario disappear while in the SP scenario they remain. For the best strategies in each scenario, both the number of queries and the average scan lengths increase as time allowance grows. Respectively, in the worst sessions for the PC case, the number of queries stays the same but the scan lengths grow as more time for search is allocated. Because the worst sessions in the PC case consume all possible queries even under the shortest time frame and the number of queries is limited, the scan lengths grow but not the number of queries. However, in the case of SP for the worst sessions, the number of queries increases and the scan length remains low as time grows. Because input costs are higher than in the PC case, investing the effort for the costly query input defines the worst behavior.

Ultimately, if there is enough time for searching, posing two or more word queries followed by a longer scan seems to provide reasonable effectiveness no matter what the search strategy among S2 to S5 is.

The third RQ is about the stability of observations. When limited time is available for searching, there is a trade-off between two action types, namely posing queries and scanning the search results cost-consciously. Thereby, the overall cost levels related to the stringency of the time frame and the relations between cost types play a major role in selecting the action type. Search interfaces and devices on which searching takes place certainly affect these variables (Kamvar & Baluja, 2007). To sum up, expensive input costs cause lengthy scanning of the search results, whereas cheap input costs help to pose better or rather longer queries. Among the QM strategies S2-S5, there is no significant difference when enough time for the search process is allocated.

Regarding the methodological RQ about the proper evaluation of search sessions under time constraints, we can state that the typical IR evaluation metrics must be applied with great care because they may be insufficient or even misleading, because traditional rank-based IR metrics do not take the user’s experience, observed costs and session gains into account. When search costs and time expended during a search session, are taken into consideration and metrics utilize normalization, i.e., scaling the value of measurement to a predefined range such as [0, 1], traditional metrics such as MAP and NDCG deliver deceptive results. Moreover, we pointed out the inappropriate use of all normalized rank-based measures.

5.4 Study IV: Modeling Behavioral Factors in Interactive Information Retrieval

In this study, we carry forward our simulation efforts with more fine-grained subtasks and more elaborate behavioral factors. As real life information access is session-based (Jansen et al., 2009), and every session consists of one or more query iterations, sessions are bound by several subtasks like query formulation[5], result scanning, document link clicking, document reading and judgment, and stopping the session. As a result, the effects of behavioral factors associated with these subtasks are inevitable. These factors include search goals and cost constraints, query formulation strategies, scanning and stopping strategies, and relevance assessment behavior, among others. The purpose of Study IV is to assess the effects of these behavioral factors on retrieval effectiveness. Our research questions include:

RQ 1: How effective is ideal human behavior, i.e., persistent scanning and ideal assessments, employing prototypical query formulation strategies, compared to deterministic baselines under various CG goals and time constraints?

RQ 2: How effective is fallible human behavior, i.e., probabilistic scanning and fallible assessments, employing prototypical query formulation strategies, compared to stochastic baselines under various CG goals and time constraints?

RQ 3: How much does fallible behavior lose in session effectiveness compared to ideal human behavior?

RQ 4: When examining the best possible query formulation strategies, is there a winning query formulation strategy which delivers the best gain across topics?

RQ 5: Methodologically, how does one simulate a behavioral model based on comprehensive session subtasks, fallible human behavior and various query formulation strategies?

In this study, we simulated both ideal human search behavior and the more realistic fallible human search behavior in an environment based on a test collection with graded relevance assessments. Our session models allowed us to simulate multiple query sessions and several interactive subtasks. During simulation experiments, the interface properties, the test collection and the search engine were kept constant and fixed probability distributions for snippet and document relevance

---

[5] Query formulation (QF in Study IV) and Query modification (QM in Study III) are interchangeably used.

assessment and for snippet scanning behaviors were utilized. Then, the following behavior factors, the use of QF strategies, cost constraints and gain goals, were varied systematically. We compared the empirically grounded prototypical QF strategies to three baselines: one long query, which comprises all available query words, with a long scanning of search results, the best possible three query session, and randomly selected QF strategy with three queries.

The first RQ was about the effectiveness of ideal human behavior employing empirically grounded QF strategies in comparison to deterministic baselines under various CG goals and time constraints. Amongst others, we found that some of empirically grounded QF strategies, second word variations (S2) and third word variations (S3) with ideal behavior are the most effective under several time constraints. They are clearly more effective than the expected effectiveness of random query sessions with ideal behavior under open time constraints in binary and non-binary scoring schemes (assigning more weights to more relevant documents), but also perform poorly compared to “one long query” sessions.

Regarding the second RQ, instead of ideal human behavior we simulated fallible human behavior with probability distributions, which were motivated by the literature (Turpin et al., 2009; Vakkari & Sormunen, 2004). Because of the random decisions based on probability distributions, simulation experiments were repeated one thousand times in order to obtain stable statements about the underlying phenomena. This approach is obviously an example of the Monte Carlo simulation method (see section 3.2.1). Again, the third word variation strategy (S3) with fallible behavior was the most effective under time constraints and gain goals. This strategy exceeds the expected effectiveness of random query sessions with fallible behavior under open time constraints for both scoring schemes, but as in the ideal case it is inferior to “one long query” sessions.

The third RQ was about the difference in effectiveness between ideal and fallible human behavior. It is no surprise that fallibility in relevance assessment and scanning decisions affected the effectiveness of sessions negatively but less with regard to highly relevant documents, because of fewer errors in their assessments. This corresponds to human selective capability and effectiveness. All in all, the effectiveness of fallible behavior is 28% to 44% of the ideal behavior.

The fourth RQ was about identifying a winning query formulation strategy across all topics. Unfortunately, there was no optimal query formulation strategy among the almost 28 000 inspected for more than one topic. Furthermore, we analyzed the query formulation strategies across the topics which perform reasonably closely, that is, within 10% of the effectiveness of the optimal strategy that is obviously distinct for every topic. Study IV shows that good session effectiveness requires a topic-focused interaction. Therefore, topics play a major role in explaining IR effectiveness. If query words are available at the beginning of a search session, a long query with persistent scanning achieves quite competitive results. However, real life searchers learn keywords from snippets seen and documents found (Ruthven, 2008). Therefore, many query words are not necessarily available at the beginning of a search session.

Regarding the fifth RQ, we employed a comprehensive multiple query session model including several subtasks, behavioral factors, goals, and constraints. We employed multiple query sessions, whereas prior simulations of interactive IR have concentrated on single query sessions (Smucker, 2009). Further, we experimented with multiple gain goals and time constraints while other studies have had limited goal and time constraints in experiment settings (Azzopardi, 2011). Moreover, we defined a snippet scanning model, which takes not only the current session gain but also the session goal and frustration explicitly into account, whereas Carterette et. al. (2011), for example, only focus on the utility gained by a single query. While we performed an exhaustive search for the best-performing strategies among all possible query formulation strategies, which can be produced under the three query and 5-keyword constraints, we also paid special attention to query formulation strategies observed in real life and analyzed their effectiveness compared to the best-performing ones. These constraints were selected, both because they reflect the typical real life interactive IR sessions and because we were limited by our computing capacity. Having more queries per session and more keywords per query would increase the required computing power exponentially (by around several orders of magnitude).

5.5 Summary of Findings

The main research questions and main findings for respective Studies I-IV are summarized in Table 4.

Table 4. Summary of main research questions and main findings

Study	Main Research Questions	Main Findings
Study I	Given RF on the first result page of PRF, is it possible to learn effective classifiers for the subsequent results?	The proposed classification approach can be applied effectively on top of PRF results. Term space reduction is not necessary.
Study II	How does RF affect IR performance when short initial queries are employed and fallible feedback is provided?	Increased error level of providing RF decreases the performance compared to perfect RF. RF with a realistic level of fallibility is as effective as perfect RF and is able to improve the performance of short initial queries.
Study III	How do various interface devices and diverse query formulation strategies affect IR sessions under overall time constraints? What is the proper evaluation methodology when time is taken into account?	If there is enough time for searching, posing two-word or longer queries followed by a longer scan seems to provide reasonable effectiveness no matter what the search strategy is. Typical rank-based IR metrics such as MAP or NDCG should be applied with great care. These metrics evaluate rankings but not user effort or experience.
Study IV	What kind and how effective are the optimal sessions under varying goals and constraints provided that human stochastic behavior is regarded?	Empirically grounded query formulation strategies, second word variations and third word variations are the most effective under several time constraints but also perform poorly compared to "one long query" sessions. Fallible behavior affects IR effectiveness negatively but less when regarding highly relevant documents because of fewer errors in their assessments.

6. Discussion and Conclusions

With “Never stop questioning”[6] as our motto, we started to question the relevance feedback and user behavior-related issues in interactive information retrieval. First, we focused on the development of novel approaches for applying relevance feedback. Namely, we utilized various standard classification and term space reduction methods in order to classify retrieved documents according to simulated user relevance feedback. As a result, correct relevance feedback was taken for granted. However, in real search environments users may very well err when they make relevance decisions based on result lists. The fallible behavior of searchers has been observed in empirical studies (Turpin et al., 2009; Vakkari & Sormunen, 2004). Nevertheless, until now experiments have been conducted with the assumption of perfect relevance assessments. To address this, we introduced fallibility in relevance feedback by defining diverse fallibility levels according to which users supply relevance feedback information to a system. Thereafter, we concentrated our research efforts on user behavior aspects, such as search behavior on different devices, which lend themselves to situational requirements, under time and search goal constraints. In this vein we brought “time” into the evaluation, which unveils some very intricate problematic points in IIR evaluation. Thus, we discovered that highly popular rank-based evaluation metrics, such as MAP and NDCG, are inappropriate for the comparison of systems when a searcher’s time expenditure is taken into account. When the search time expended by users is part of the evaluation process, normalized rank-based metrics may provide misleading evaluation results. Therefore, non-normalized metrics should be employed. Finally, we elaborated our simulation experiments by defining fine-grained user behavior variables. For example, search strategies, search goals and cost constraints, scanning and assessment behavior, and relevance scoring were incorporated into the design of simulation experiments. We applied both deterministic and stochastic approaches

[6] Albert Einstein’s quote

during the simulation of searcher behavior and contrasted both approaches to narrow the gap between traditional Cranfield-type experiments and real life search behavior. Moreover, user behavior is always assumed correctly in Cranfield-type experiments. However, in our experiments, we adulterated the simulated user behavior with some probabilities which were set according to prior empirical studies (Turpin et al., 2009; Vakkari & Sormunen, 2004) to reflect the real user's interaction with a search system.

Table 5. Summary of themes and variables of the contributed and Cranfield-style studies. Variables are encoded as: fixed variables are lowercase, independent (varied) variables are uppercase and dependent variables are bold and uppercase.

Study	**Theme**	**Subtheme**	**Signatures of Variables**
Cranfield-style	Various	Various	V M q - - - - - **E**
Study I	RF	Classification	V M q t s a c j **E**
Study II	RF	Fallibility	V m Q t s A c J **E**
Study III	Sessions	QF/Time/Interfaces	V m Q T S a c j **E**
Study IV	Sessions	QF/Frustration/Patterns	V m Q T S A C J **E**

In order to give an overview of the studies, the main and sub-themes together with the variables of Studies I to IV are summarized in Table 5. The relevant variables of each study are encoded in the table as follows: variables which are fixed during experiment execution are given in lowercase, independent variables which are varied, are denoted in uppercase, and dependent variables which are examined, are bolded. In Table 5, *v* stands for vocabulary, and consequently represents information need modeling, *m* stands for the retrieval method applied, *q* for querying, *t* for time consumed, *s* for scanning the search results, *a* for assessing the relevant snippets or documents, *c* for clicking the document links, *j* for judging the document relevance, and *E* for effectiveness of retrieval in the IR experiments. Accordingly, in order to analyze the new approach to RF, we fixed all interaction variables except the method (*M*) in Study I. In Study II we varied the user's judgment of document relevance, and explored its effects on performance. Further, in Study III not only the scanning patterns but also time is varied and effectiveness is analyzed. Finally, in Study IV we varied time and scanning patterns as well as assessing, clicking and judging behavior in order to examine the effectiveness. In comparison, Cranfield-style experiments do not take the variables into account at all. These are depicted with a dash in Table 5. In Studies I to IV we progressively introduced and varied such variables, and furthermore we shed light on the effects of these variables.

In Study I, we applied standard classification algorithms to select the relevant documents from the result list, more precisely from the second page onwards, after relevance feedback was supplied on the first page of the search engine results. Thereby, we simulated the relevance feedback on the first page of results of the pseudo-relevance feedback run of topical queries. Thus, classifiers were able to learn relevant and non-relevant document classes, which are used to decide on the relevance of further documents on the result list. Moreover, we explored numerous term space reduction techniques (Sebastiani, 2002) for improving both effectiveness and efficiency of the classification process. Comprehensive experiments on TREC ad-hoc collections (Voorhees & Harman, 2000) indicated that this approach of applying RF with the help of classification methods is significantly more effective than PRF with title queries as well as title-and-description queries. However, the difference between classification approaches and the combination of classification methods with term space reduction methods entail no significant improvement from the statistical view point. In other words, any of the state-of-the-art classification methods can improve the PRF results of short queries. However, in order to learn a classifier, the first page of results should have at least one relevant and one non-relevant document, otherwise the necessity of inspecting the following pages may disappear because of the correlation of the first and subsequent result pages. When the first page has a lot of relevant documents, the user's information need may already be satisfied (Sakai, 2006), and vice versa, when the first page has no relevant document, the following pages are not likely to have enough relevant documents worth being classified or inspected either.

One could try to infer RF automatically e.g., from users' interaction with search interfaces (Ruthven, 2008), and thereafter apply the classification approach to amend the result lists. Overall, the parallelism between our approach and Learning to Rank algorithms (Li, 2011) should be emphasized once more while regarding how differently both approaches are applied to improving search results.

In Study II, we simulated the RF further but this time we suggested a novel approach to RF evaluation, i.e., we introduced and systematically studied the effects of the searchers' fallibility in supplying RF. This can occur in real life search situations, because the snippets delivered by a search engine could mislead searchers into deciding incorrectly about the relevance of a document (Turpin et al., 2009). Further, the multi-grade relevance assessments (Sormunen, 2002) were employed in this study to improve the realism of simulations by regarding the expertise levels of searchers. Moreover, the experiments were carried out with the short initial queries generated by real searchers (Keskustalo et al., 2009) instead of the longer title-and-description queries, which are quite popular in Cranfield-style IR experiments. Our findings indicated that the results of very short initial queries can be improved by applying query-biased summaries (Bates, 1989; Tombros & Sanderson, 1998; Turpin et al., 2009) for both PRF and direct user-supplied RF. Furthermore, the experimental results showed that very fallible feedback is no different from pseudo-relevance feedback (PRF) and not effective in short initial queries. However, RF with empirical fallibility is practically as effective as correct RF and is able to improve the performance of short initial queries. RF systematically improves the performance of all short-query types when evaluation is liberal; but does not improve against PRF when evaluation is strict. Short initial queries obviously do not provide enough good documents for strict evaluation. It is not surprising that in real life users prefer to revise their queries instead.

In Study III our attention was drawn to session simulations away from the RF simulations. Real life search experience is characterized by time constraints, multiple short queries and a myriad of different search interfaces on different devices. Because busy life situations require prompt answers to ad-hoc emerging questions, ubiquitous computing lends itself to such kinds of needs very well. However, different devices demand different interaction styles; consequently time plays a major role during query input phase. First, we explored the space for all possible multi-query sessions within the limited framework for the best querying and scanning behavior under diverse time and goal constraints. Interestingly, there was no single winning query modification pattern which performed best across all the topic queries. Both the number of queries and the average scan lengths increase in the best strategies for both PC and SP scenarios analyzed when time allocation grows, whereas the number of queries does not change but scan length grows in the worst sessions for personal computer scenario, because the worst sessions consume all the available queries as soon as possible. On the other hand, for the smartphone scenario, while the number of queries grows, the scan lengths remain low when more time is available for searching. Because the input costs are high, investing the effort in query typing is apparently the worst behavior.

The typing speed supported by a device or an interface determines a user's input effort. Where the input effort is low, for example in the PC case, better or rather longer queries are favorable. However, if the input effort is high, such as in the SP case, users are not able to type longer queries under time pressure. Instead they must perform lengthy scans of weak short-query results in order to achieve their goals. When the user has enough time at their disposal, the way in which searches are executed is not crucial, because there will be enough time to identify the relevant documents.

Finally we discovered the peculiarities of time-based evaluation with standard rank-based evaluation metrics. Typical IR evaluation metrics (Demartini & Mizzaro, 2006; Su, 1992) are based on the quality of ranking alone. When time is taken into account, normalized metrics may deliver misleading results, and therefore should be used with care. Furthermore, in session-based evaluation they must also be applied with great care because they may be insufficient or even misleading. They may be partially insensitive to the user's experience, e.g., because of recurrent documents, and observed costs and benefits. This is particularly critical when a user's costs (time expenditure) are taken into account and the metric employs normalization, i.e., scaling the measurements to a predefined range such as [0, 1]. For example, the popular NDCG metric (Järvelin & Kekäläinen, 2002) and its non-discounted counterpart NCG should be avoided in any comparisons between searching environments, and between strategies within a given searching environment when user effort is taken into account. This is because the ideal gain vector used for normalization is read to vastly different lengths between strategies or environments. For example, consider Figure 4, which plots NCG over time for QM strategy S2 in two scenarios, PC and SP, from Study III.

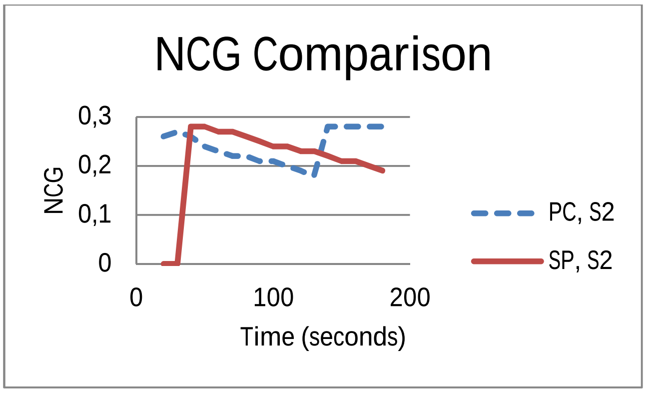

Figure 4. NCG vs. time comparison of PC and SP scenarios for QM strategy S2

Due to normalization (division by the ideal cumulated gain vector), the smartphone (SP) scenario seems to exhibit better performance in the time frame from 40 to 135 seconds. This is due to (a) ranking being somewhat effective, and (b) the number of documents seen in each session: in the PC case the user sees 15 to 35 documents, but in the SP case only 5 to 20 documents in the indicated time frame. Figure 5 plots CG with the corresponding data and makes the difference clear. Similar pitfalls also plague the most classic metric, MAP (see Chapter 6 in Study III).

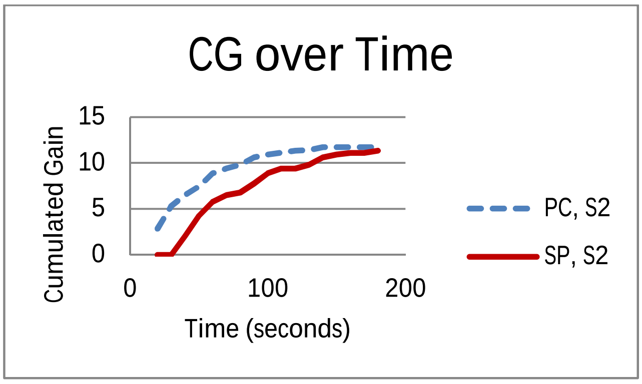

Figure 5. CG over time for QM strategy S2 in PC and SP scenarios

Even within a non-normalized metric like CG, incorporating time in session-based evaluation has profound effects. Consider figures 6 and 7. The former gives traditional cumulated gain over ranks for QM strategies S1 and S3 averaged over the topics. The latter gives CG over time for the same strategies in the two scenarios.

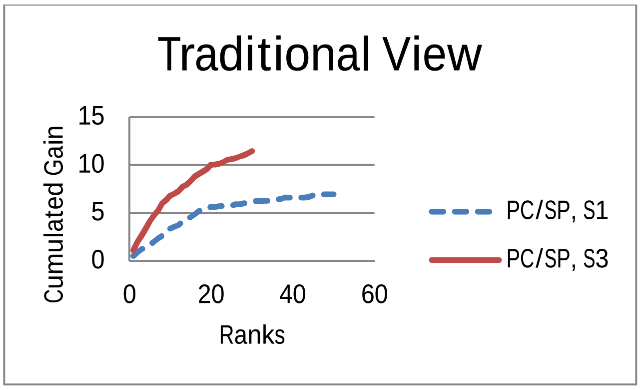

Figure 6. Traditional View, CGs over ranks, scenarios PC and SP for QM strategies S1 (allowing five queries) and S3 (allowing only three queries)

In Figure 6, both the PC and SP scenarios have the same observed effectiveness, because the evaluation focuses on the gain (CG) over the result ranks, regardless of how long it takes to retrieve the documents. The two strategies S1 and S3 differ in effectiveness, S3 providing far better effectiveness than S1. However, when time is taken into account (Fig. 7), the scenarios and strategies differ greatly from each other. Up to 60 seconds, S3 in the SP case is the worst strategy and this is entirely due to the high input cost of the long query. With enough time (180 sec.), S3 in SP catches up with S3 in the PC case. Also, PC and SP do not much differ for S1 due to the relatively low input cost and the weak result quality. Comparing figures 6 and 7, it is easy to see that time profoundly affects both user experience and effectiveness in sessions in different scenarios.

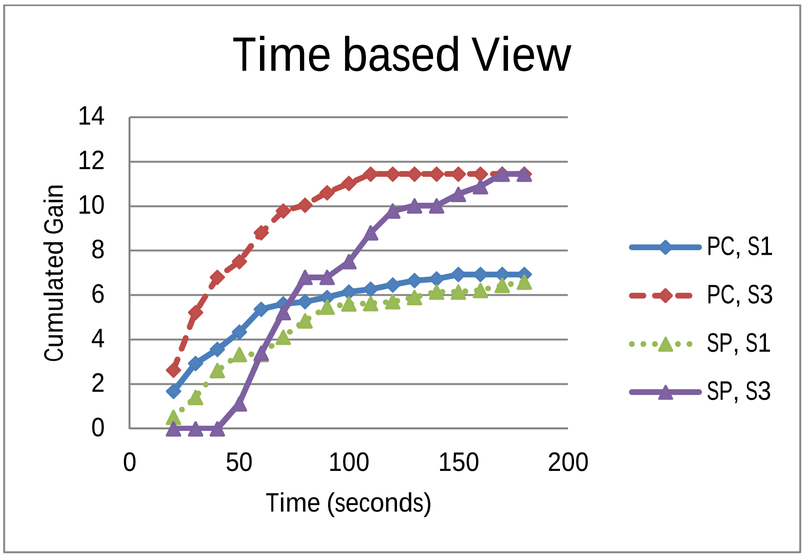

Figure 7. Time-based View, CGs over time, scenarios PC and SP for strategies S1 and S3

In terms of time, we employed cumulated gain in our experiments in Study III. But in order to compare various devices and search interfaces with respect to time-based evaluation, especially in different studies, the normalization issue remains to be addressed and is an interesting study subject.

In Study IV we delved into user behavior issues during query result inspection. We modeled full multi-query sessions with comprehensive subtasks. Thus we extended the previously defined user models (Keskustalo, 2010) with further details about scanning the snippets, deciding on clicking the links, reading the documents, and judging the relevance of documents. Moreover, we conducted experiments based on both deterministic and stochastic behavior. While the deterministic approach dictates certain predictable behavior, the stochastic case requires probabilities about the outcome of every decision. Therefore, we defined probability distributions beginning with the deterministic case and moving towards a totally random one. Then, we compared the prototypical query modification strategies with the best possible strategy. Hence, we can further suggest the prototypical strategies, especially second and third word variations (S2 and S3), for future research activities, because even they do not reach the level of best performer patterns; they still represent a regular pattern and are on par with best performers to some degree. In the ideal case, S3 achieves about three-quarters of the performance of the long query and of the by-topic optimized best session pattern, which means the strategy that is distinctly optimized for each topic. All prototypical QM strategies except the sequence of one-word queries (S1) are close to each other in terms of effectiveness, with both ideal and fallible behavior. Among all possible QM strategies inspected, around 28 000, there was not one strategy that was best across all the topics. Further analysis showed that the next to best-performing strategies, the effectiveness of which is at least 90% of the best strategy for each particular topic, across all the topics achieve good performance in only around one-third of the topics. This advocates the view that users apply topic-specific QM strategies in order to reach the highest possible effectiveness.

Because the results of stochastic experiments depend on the selected probability values, the experiments are repeated in order to get an average value over a wide range of possible values. In conclusion, stochastic behavior was obviously inadequate in comparison with deterministic one, which is not realistic although it is superior. Probabilistic scanning and fallible relevance assessment limit the performance but cause considerably less damage regarding highly relevant documents. Another analysis showed that single long queries yield better performance levels than multiple query sessions. However, in real life searchers do learn during interaction with search results, frequently modify queries accordingly (Ruthven, 2008) and do not initially have a long query available.

Methodologically we extend the use of traditional test collections to include behavioral factors in a controlled experimental environment in order to study the effects of searcher-related factors in IIR. Furthermore, we simulated the behavioral factors using the Monte Carlo method (Altiok & Melamed, 2007, pp. 11-22) based on behaviors observed in real life studies (Vakkari & Sormunen, 2004). We also integrated the time variable into the evaluation process. Moreover, we simulated user interactive sessions on different interfaces. Besides, we examined all possible query formulation patterns created by a limited vocabulary in order to analyze their effects on IIR effectiveness.

In general, with every article the researchers learn new concepts and gain insight into diverse phenomena. In retrospect, so did we. Thus the recently gained knowledge can now be fed into the research settings of the previous studies as well as blended with new research topics; thereby very intricate research problems may arise, such as searcher behaviors in several stages of interaction with various search systems and environments as well as in the evaluation process.

However, one should bear in mind that the applied methods and findings are not limited to topical search, but they can be exploited across a wider field of information retrieval. When the importance of information retrieval in the current and future knowledge society is considered, the contribution of bringing the searcher’s behavioral aspects into the fast-paced IR experimental world can be better appreciated.

With this thesis and its contribution to the IR research community, we hope we are able to instigate new approaches, methods, and metrics for interactive information retrieval.

References

Alpert, J. & Hajaj, N. (2008, July 25). Official Blog: We knew the web was big. [Web log post]. Retrieved from http://googleblog.blogspot.fi/2008/07/we-knew-web-was-big.html

Altiok, T., & Melamed, B. (2007). *Simulation modeling and analysis with ARENA.* Burlington, MA, USA: Academic Press.

Azzopardi, L. (2011). The economics in interactive information retrieval. *In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval,* Beijing, China. 15-24. doi:10.1145/2009916.2009923.

Azzopardi, L., Järvelin, K., Kamps, J., & Smucker, M. D. (2011). Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum, 44*(2), 35-47. doi:10.1145/1924475.1924484

Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the ngram statistics package. *Computational linguistics and intelligent text processing* (pp. 370-381) Springer.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Information Review, 13*(5), 407-424.

Belew, R. K. (2000). *Finding out about: A cognitive perspective on search engine technology and the WWW.* Cambridge University Press.

Belkin, N. J. (1980). Anomalous states of knowledge. *Canadian Journal of Information Science, 5*, 133-143.

Belkin, N. J., & Croft, W. B. (1987). Retrieval techniques. *Annual Review of Information Science and Technology, 22*, 109-145.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human computer interaction.* Routledge.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Comput.Surv., 44*(1), 1-50. doi:10.1145/2071389.2071390

Carterette, B., Kanoulas, E., & Yilmaz, E. (2011). Simulating simple user behavior for system effectiveness evaluation. *In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management,* Glasgow, Scotland, UK. 611-620. doi:10.1145/2063576.2063668.

Catarci, T., & Kimani, S. (2013). Human-computer interaction view on information retrieval evaluation.7757, 48-75. doi:10.1007/978-3-642-36415-0_3

Chapelle, O., & Chang, Y. (2011). Yahoo! learning to rank challenge overview. *Journal of Machine Learning Research-Proceedings Track, 14*, 1-24.

Clarke, C. L. A., Freund, L., Smucker, M. D., & Yilmaz, E. (2013). Report on the SIGIR 2013 workshop on modeling user behavior for information retrieval evaluation (MUBE 2013). SIGIR Forum, 47(2), 84-95. doi:10.1145/2568388.2568403

Cleverdon, C. W., Mills, J., & Keen, M. (1966). Factors determining the performance of indexing systems. New York

Conover, W. J. (1999). *Practical nonparametric statistics.* John Wiley & Sons.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice.* Addison-Wesley Reading.

Demartini, G., & Mizzaro, S. (2006). A classification of IR effectiveness metrics.*3936*, 488-491. doi:10.1007/11735106_48

Dunlop, M. D. (1997). Time, relevance and interaction modelling for information retrieval. *SIGIR Forum, 31*(SI), 206-213. doi:10.1145/278459.258569

Efthimiadis, E. (1996). Query expansion. *Annual Review of Information Science and Technology, 31*, 121-187.

Fidel, R. (1985). Moves in online searching. *Online Information Review, 9*(1), 61-74.

Foley, C., & Smeaton, A. F. (2009). Synchronous collaborative information retrieval: Techniques and evaluation. *In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval,* Toulouse, France. 42-53. doi:10.1007/978-3-642-00958-7_7.

Harman, D. (1992). Relevance feedback revisited. *In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1-10.

Hill, T., & Lewicki, P. (2007). *STATISTICS: Methods and applications.* Tulsa, OK.: StatSoft.

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context (the information retrieval series).* Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Jansen, B. J., Booth, D. L., & Spink, A. (2009). Patterns of query reformulation during web searching. *Journal of the American Society for Information Science and Technology, 60*(7), 1358-1371. doi:10.1002/asi.21071

Jansen, B. J., & Spink, A. (2006). How are we searching the world wide web? A comparison of nine search engine transaction logs. *Information Processing & Management*, *42*(1), 248-263.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Inf.Process.Manage.*, *36*(2), 207-227. doi:10.1016/S0306-4573(99)00056-4

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. 2053-2056. doi:10.1145/1645953.1646299.

Järvelin, K., & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. *In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Athens, Greece. 41-48. doi:10.1145/345508.345545.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Trans.Inf.Syst.*, *20*(4), 422-446. doi:10.1145/582415.582418

Joachims, T. (1999). Making large scale SVM learning practical.

Joachims, T., & Radlinski, F. (2007). Search engines that learn from implicit feedback. Computer, 40(8), 34-40. doi:10.1109/MC.2007.289

Kamps, J., Geva, S., Peters, C., Sakai, T., Trotman, A., & Voorhees, E. (2009). Report on the SIGIR 2009 workshop on the future of IR evaluation. SIGIR Forum, 43(2), 13-23. doi:10.1145/1670564.1670567

Kamvar, M., & Baluja, S. (2007). Deciphering trends in mobile search. *Computer, 40*(8), 58-62. doi:10.1109/MC.2007.270

Kamvar, M., Kellar, M., Patel, R., & Xu, Y. (2009). Computers and iphones and mobile phones, oh my!: A logs-based comparison of search users on different devices. *In: Proceedings of the 18th International Conference on World Wide Web,* 801-810.

Kanoulas, E., Carterette, B., Clough, P. D., & Sanderson, M. (2011). Evaluating multi-query sessions. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 1053-1062.

Karat, C., Halverson, C., Horn, D., & Karat, J. (1999). Patterns of entry and correction in large vocabulary continuous speech recognition systems. *In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Pittsburgh, Pennsylvania, USA. 568-575. doi:10.1145/302979.303160.

Kekäläinen, J., & Järvelin, K. (2002). Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. *In: Proceedings of the 4th CoLIS Conference*, 253-270.

Kelly, D. (2005). Implicit feedback: Using behavior to infer relevance. *New directions in cognitive information retrieval* (pp. 169-186) Springer.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*(1—2), 1-224.

Keskustalo, H. (2010). Towards simulating and evaluating user interaction in information retrieval using test collections. Ph.D. thesis. Tampere, Finland: University of Tampere

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: Proceedings of the 28th European Conference on Advances in Information Retrieval,* London, UK. 191-204. doi:10.1007/11735106_18.

Keskustalo, H., Järvelin, K., & Pirkola, A. (2008). 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), 209-228. doi:10.1007/s10791-007-9043-7

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: Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology,* Sapporo, Japan. 63-74. doi:10.1007/978-3-642-04769-5_6.

Lam-Adesina, A. M., & Jones, G. J. F. (2001). Applying summarization techniques for term selection in relevance feedback. *In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* New Orleans, Louisiana, USA. 1-9. doi:10.1145/383952.383953.

Lavrenko, V., & Croft, W. B. (2001). Relevance based language models. *In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, USA. 120-127. doi:10.1145/383952.383972

Li, H. (2011). Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, *4*(1), 1-113.

Losee, R. M. (1998). *Text retrieval and filtering: Analytic models of performance.* Kluwer Academic Publishers.

Maria, A. (1997). Introduction to modeling and simulation. *In:Proceedings of the 29th Conference on Winter Simulation*, Atlanta, Georgia, USA. 7-13. doi:10.1145/268437.268440

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge University Press Cambridge.

Marchionini, G. (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), 35-69.

Marchionini, G. (1995). *Information seeking in electronic environments.* New York, NY, USA: Cambridge University Press.

Oard, D. W., & Kim, J. (2001). Modeling information content using observable behavior.

Qin, T., Liu, T., Xu, J., & Li, H. (2010). LETOR: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, *13*(4), 346-374.

Ricardo, B. Y. (1999). *Modern information retrieval.* Pearson Education.

Ruthven, I. (2008). Interactive information retrieval. *Annual Review of Information Science and Technology.*, *42*(1), 43-91. doi:10.1002/aris.144.v42:1

Ruthven, I., & Kelly, D. L. (2011). *Interactive information seeking, behaviour and retrieval.* Facet Publishing.

Ruthven, I., & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review., 18*(2), 95-145. doi:10.1017/S0269888903000638

Ruthven, I., Lalmas, M., & Van Rijsbergen, K. (2003). Incorporating user search behavior into relevance feedback. *Journal of the American Society for Information Science and Technology, 54*(6), 529-549. doi:10.1002/asi.10240

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,* Seattle, Washington, USA. 695-696. doi:10.1145/1148170.1148322.

Salton, G. (1970). Evaluation problems in interactive information retrieval. *Information Storage and Retrieval, 6*(1), 29-44.

Saracevic, T. (1996). Relevance reconsidered. *In: Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2),* 201-218.

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. *In: Proceedings of the American Society for Information Sciences, 34* 313-327.

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, 30,* 3-71.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology, 58*(13), 2126-2144.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys, 34*(1), 1-47. doi:10.1145/505282.505283

Siegel, S., & Castellan, N.J. (1988). Nonparametric statistics for the behavioral sciences. *McGraw-HiU Book Company, New York*.

Smucker, M. D. (2009). Towards timed predictions of human performance for interactive information retrieval evaluation. *In: Third Workshop on Human-Computer Interaction and Information Retrieval (HCIR'09).*

Smucker, M. D., & Clarke, C. L. A. (2012a). Time-based calibration of effectiveness measures. *In:Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval,* Portland, Oregon, USA. 95-104. doi:10.1145/2348283.2348300

Smucker, M. D., & Clarke, C. L. A. (2012b). Stochastic simulation of time-biased gain. *In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management,* Maui, Hawaii, USA. 2040-2044. doi:10.1145/2396761.2398568.

Smucker, M. D., & Clarke, C. L. A. (2012c). Modeling user variance in time-biased gain. *In: Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval,* Cambridge, California. 3 1-10. doi:10.1145/2391224.2391227.

Sokolowski, J. A., & Banks, C. M. (Eds.). (2011). *Principles of modeling and simulation: a multidisciplinary approach*. John Wiley & SonsSormunen, E. (2002). Liberal relevance criteria of TREC -: Counting on negligible documents? *In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Tampere, Finland. 324-330. doi:10.1145/564376.564433.

Strohman, T., Metzler, D., Turtle, H., & Croft, W. B. (2005). Indri: A language model-based search engine for complex queries. *In: Proceedings of the International Conference on Intelligent Analysis, 2*(6).

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing & Management, 28*(4), 503-516. doi:10.1016/0306-4573(92)90007-M

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. *In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Melbourne, Australia. 2-10. doi:10.1145/290941.290947.

Toms, E. (2013). User-oriented information retrieval.7757, 76-85. doi:10.1007/978-3-642-36415-0_4

Turpin, A., Scholer, F., Jarvelin, K., Wu, M., & Culpepper, J. S. (2009). Including summaries in system evaluation. *In: Proceedings of the 32nd International ACM SIGIR*

Conference on Research and Development in Information Retrieval, Boston, MA, USA. 508-515. doi:10.1145/1571941.1572029.

Vakkari, P. (2000). Cognition and changes of search terms and tactics during task performance: A longitudinal case study. *In: RIAO,* 894-907.

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology, 37*(1), 413-464. doi:10.1002/aris.1440370110

Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation, 56*(5), 540-562.

Vakkari, P., & Sormunen, E. (2004). The influence of relevance levels on the effectiveness of interactive information retrieval. *Journal of the American Society for Information Science and Technology, 55*(11), 963-969. doi:10.1002/asi.20046

Voorhees, E. M., & Harman, D. K. (2000). The eigth text REtrieval conference (TREC-8), volume 8. national institute of standards and technology, NIST. *NIST Special Publication,* 500-246.

White, R. W. (2011). Interactive techniques. In I. Ruthven, & D. Kelly (Eds.), *Interactive information seeking, behaviour and retrieval* (pp. 171) Facet Publishing.

White, R. W., & Drucker, S. M. (2007). Investigating behavioral variability in web search. *In: Proceedings of the 16th International Conference on World Wide Web,* 21-30.

White, R. W., Ruthven, I., & Jose, J. M. (2002). The use of implicit evidence for relevance feedback in web retrieval.*2291*, 93-109. doi:10.1007/3-540-45886-7_7

Yi, J., Maghoul, F., & Pedersen, J. (2008). Deciphering mobile search patterns: A study of yahoo! mobile search queries. *In: Proceedings of the 17th International Conference on World Wide Web,* Beijing, China. 257-266. doi:10.1145/1367497.1367533.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. *In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 334-342.

Appendix

Term Space Reduction algorithms

In order to utilize various term space reduction methods (see for the formulas: Banerjee & Pedersen, 2003; Sebastiani, 2002; Siegel & Castellan, 1988, pp. 102-166 & 224-254), for each term, a 2 x 2 matrix is created, which incorporates the number of relevant and non-relevant documents in which a respective term occurs or does not occur (see Table A1). With the help of this matrix, a particular measure for the respective term can be calculated according to the formulas given below. Having a list of terms ordered by the magnitude of the calculated measure, the number of possible terms can be pruned to the desired size by neglecting the terms that are less significant.

Table A1. Term matrix for term space reduction

# of documents	Relevant Documents	Non-relevant Documents
Term existence	n_{11}	n_{12}
Term absence	n_{21}	n_{22}

In the table, n_{11} is the number of relevant documents in which the current term occurs; n_{12} is the number of non-relevant documents in which the current term occurs; n_{21} is the number of relevant documents in which the current term does not occur; and n_{22} is the number of non-relevant documents in which the current term does not occur.

Mutual information gain (MIG) measures the mutual dependence of two random variables. MIG is the expected value of pointwise mutual information. The formula for MIG is given below. The expected value (E_{ij}) for the pertinent cell position is the ratio of the product of marginal to the total number of frequencies, e.g., the number of documents:

$$E_{ij} = \frac{(\sum_{column} O) \cdot (\sum_{row} O)}{\sum O}, in\ first\ cell: \mathrm{E}_{11} = \frac{(n_{11} + n_{21}) \cdot (n_{11} + n_{12})}{(n_{11} + n_{12} + n_{21} + n_{22})}$$

Where O is the observed value for the respective cell.

$$\mathrm{MIG} = \sum_{i} \sum_{j} n_{ij} \cdot \log\left(\frac{n_{ij}}{E_{ij}}\right)$$

Pearson's chi-squared test examines the variables according to chi-squared test. The test value will be calculated by summing the normalized squared deviations between observed and theoretically expected frequency.

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where X^2 is the Pearson test statistics, O_i is observed frequency, E_i is expected frequency, n is the number of cells in the tables, i.e., 4.

Odds ratio also measures the association of two variables.

$$\text{Odds ratio} = \frac{(n_{11} \cdot n_{22})}{(n_{21} \cdot n_{12})}$$

Kendall-Tau rank correlation coefficient, as the name suggests, measures the rank correlation of two variables. It describes the similarity of orderings of variables, and can be calculated thus:

$$\tau = \frac{\#\ of\ concordant\ pairs - \#of\ discordant\ pairs}{0.5 * n \cdot (n - 1)}$$

Where n is the number of observations. Any pair of observations of two variables e.g., (x_i,y_i), (x_j y_j), are concordant if both values of the pair are either greater (x_i>x_j and y_i>y_j) or smaller (x_i<x_j and y_i<y_j) than in the pair. Otherwise, the pairs are discordant unless the pair values are the same. The contingency table can be expressed as observation values of two variables in order to build concordant and discordant pairs. Then the Kendall-Tau rank correlation formula can be calculated.

The Spearman rank correlation coefficient is a non-parametric measure like the Kendall-Tau rank correlation, which examines the dependence of two variables.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Where x_i and $y_i$represent sample vectors, and $\bar{x}\ and\ \bar{y}$ are sample means. Again, the contingency table can be reformulated as x and y vectors, e.g., x = [n_{11}, n_{12}] and y= [n_{21}, n_{22}], thereafter the Spearman rank correlation formula can be applied.

Fisher's exact test is an exact test to define the association of two variables in a contingency table; in the current thesis these variables are the existence and absence of terms. The probability of obtaining exactly these values as in contingency tables is given by hyper-geometric distribution. Fisher's exact test formula is given as:

$$p = \frac{\binom{n_{11} + n_{12}}{n_{11}} \cdot \binom{n_{21} + n_{22}}{n_{21}}}{\binom{n}{n_{11} + n_{21}}}$$

Where n is the sum of all cell values, and $\binom{a}{b}$ represents the binomial coefficient, which is calculated as $\frac{a!}{b!(a-b)!}$ and gives the number of 'b' element subsets from 'a' elements. As a result, the first binomial component of the numerator calculates the combinatorial number of term existence in relevant documents, and the second multiplier calculates the combinatorial number of term absence in non-relevant documents. Multiplication of both these numbers gives the number of all possible combinations, which is divided by the number of selecting all possible combinations of relevant documents in the document collection. Finally, the result of this division, which denotes the relative frequency of the occurrence of an experiments outcome, results in the probability of the term occurrence in relevant documents. Therefore, the probabilities of terms can be compared further, so as to order them accordingly.