DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
81

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
81

GEORG SINGER

Web search engines and complex information needs

UNIVERSITY OF TARTU
PRESS
1632

Institute of Computer Science, Faculty of Mathematics and Computer Science, University of Tartu, Estonia

This dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Informatics on 19 October, 2012, by the Council of the Institute of Computer Science, Faculty of Mathematics and Computer Science, University of Tartu.

Supervisors:

Prof. Eero Vainikko Institute of Computer Science University of Tartu Tartu, Estonia	Dr. Ulrich Norbisrath Institute of Computer Science University of Tartu Tartu, Estonia	Prof. Dirk Lewandowski Department of Information Hamburg University of Applied Sciences Hamburg, Germany

Opponents:

Prof. René Schneider Haute Ecole de gestion Dép. Information documentaire Rte de Drize 7, 1227 Carouge, Switzerland	Prof. Joachim Griesbaum Stiftung Universität Hildesheim Lübecker Straße 3 31141 Hildesheim, Germany

Commencement will take place on 19 October, 2012, at 16.15 at Liivi 2 - 405, 50409, Tartu, Estonia.

ISSN 1024-4212
ISBN 978-9949-32-110-0 (print)
ISBN 978-9949-32-111-7 (pdf)

University of Tartu Press
www.tyk.ee
Order No. 446

Contents

List of original publications

I. Singer, G.[1]; Danilov, D.[1]; Norbisrath, U. (2012). Complex Search: Aggregation, Discovery, and Synthesis. *Proceedings of the Estonian Academy of Sciences*, 61, 2, pp. 89-106

II. Singer, G.; Norbisrath, U.; Vainikko, E.; Kikkas, H.; Lewandowski, D. (2011). Search-Logger - Analyzing Exploratory Search Tasks. *Proceedings of the 2011 ACM Symposium on Applied Computing,* SAC '11, Taiwan, ACM New York, pp. 751-756

III. Singer, G.; Norbisrath, U.; Lewandowski, D. (2012). An Experiment with Ordinary Web Search Engine Users Carrying Out Complex Search Tasks. Manuscript submitted for publication.

IV. Singer, K.[1]; Singer, G.[1]; Lepik, K.; Norbisrath, U.; Pruulmann-Vengerfeldt, P. (2011). Search Strategies of Library Search Experts. *3rd International Conference on Qualitative and Quantitative Methods in Libraries*, QQML 2011, Athens, Greece, in press.

V. Singer, G.; Pruhlmann-Vengerfeldt, P.; Norbisrath, U.; Lewandowski, D. (2012). The Relationship between Internet User Type and User Performance when Carrying out Simple vs. Complex Search Tasks. *First Monday*, 17, 6.

VI. Singer, G.; Norbisrath, U.; Lewandowski, D. (2012). Ordinary Web Search Engine Users assessing Difficulty, Effort, and Outcome for Simple and Complex Search Tasks. *4th Information Interaction in Context Symposium iiiX 2012*, Nijmegen, The Netherlands, in press.

VII. Singer, G.; Norbisrath, U.; Lewandowski, D. (2012). Impact of Gender and Age on Performing Search Tasks Online. *Mensch und Computer* 2012, Konstanz, Germany, in press.

Other non related publications

- Livenson, I.; Singer, G.; Srirama, S.N.; Norbisrath, U.; Dumas, M.; (2010). Towards a Model for Cloud Computing Cost Estimation with Reserved Resources. *Proceedings of the 2nd International ICST Conference on Cloud Computing (CloudComp 2010).*

[1] Equal contribution

- Kirsimäe, S.; Norbisrath, U.; Singer, G.; Srirama, S.; Lind, A. (2011). Extending Friend-to-Friend Computing to Mobile Environments. *Proceedings of the First International Conference on Mobile Services, Resources, and Users*; Barcelona, Spain; pp. 75 - 80 - best paper award

- Tamme, T.[1]; Norbisrath, U.[1]; Singer, G.; Vainikko, E. (2011). Improving Email Management. *Proceedings of the First International Conference on Advances in Information Mining and Management*; Barcelona, IMMM 2011, Barcelona, Spain, pp. 67 - 72.

Abstract

Search engines have become the means for searching information on the Internet. Along with the increasing popularity of these search tools, the areas of their application have grown from simple look-up to rather complex information needs. Also the academic interest in search has started to shift from analyzing simple query and response patterns to examining more sophisticated activities covering longer time spans. Current search tools do not support those activities as well as they do in the case of simple look-up tasks. Especially the support for aggregating search results from multiple search-queries, taking into account discoveries made and synthesizing them into a newly compiled document is only at the beginning and motivates researchers to develop new tools for supporting those information seeking tasks.

In this dissertation I present the results of empirical research with the focus on evaluating search engines and developing a theoretical model of the complex search process that can be used to better support this special kind of search with existing search tools[2].

I present a model that decomposes complex Web search tasks into a measurable, three-step process. I show the innate characteristics of complex search tasks that make them distinguishable from their less complex counterparts and showcase an experimentation method to carry out complex search related user studies. I demonstrate the main steps taken during the development and implementation of the Search-Logger study framework (the technical manifestation of the aforementioned method) to carry our search user studies. I present the results of user studies carried out with this approach. Finally I present development and application of the ATMS (awareness-task-monitor-share) model to improve the support for complex search needs in current Web search engines.

[2]It is not the goal of the thesis to implement a new search technology. Therefore performance benchmarks against established systems such as question answering systems are not part of this thesis.

Chapter 1

Introduction

> "The ultimate search engine would basically understand everything in the world, and it would always give you the right thing. And we're a long, long ways from that." Larry Page - Google Founder [1]

According to Maslow's hierarchy of needs [81] information seeking is a fundamental human activity. Searching for necessary information is more and more shifting to electronic media like the Internet [40]. Since the invention of the Web and its rapid growth over the last decade, the amount of information available on this medium has become overwhelming. Information overload imposes a growing problem upon our knowledge societies, impacting productivity at the workplace level and also influencing the end user Internet experience [17, 64, 42]. Search engines are the means to search for information on the Internet. People use search engines for all kinds of tasks, from simply looking up facts to planning their holiday trips and their investment decisions. While looking up facts is well supported by current Web search engines [65, 77], this does not hold true in case of more complex search tasks [115], where an increasing number of users is dissatisfied [113].

For further reference and imagination, I introduce the following scenario (based on [98]), illustrating a complex search task (I will define the concept of a complex search task in Chapter 2 on page 17):

> Imagine Brian and Sarah are journalists in the popular newspaper FooTimes. Brian and Sarah are currently working on a political analysis article related to the conflict between North Korea and South Korea. They decide to divide the work. Brian will search for related information and events that happened in North Korea dur-

ing the last three-month period, while Sarah will search for events and facts, which took place in South Korea.

In the given case, the journalists need to have many search sessions, store the relevant articles and information sources, combine them, and research relations between different events.

First, the journalists search for reliable information sources. They start from searching for articles on major news web portals like BBC, CNN, or Financial Times. Brian and Sarah also use search engines and social bookmarking services like Delicious (see e.g. [41]) to search for local newspapers and other public information sources like information on embassy Web sites. During the search Brian and Sarah quickly examine the information on the Web pages and create bookmarks for the relevant ones. They take notes to remember the most important facts to later find relations between the information in the relevant documents.

This complex search goes far beyond a simple look-up task and includes discovery, aggregation, and synthesis tasks. As we see the task of searching different Websites alone and then picking the right information for later research is complex. It means keeping track of the sources, synthesizing relevant information and sharing with co-workers. Finally it is essential to identify relations between the information and draw conclusions from that.

Today's search systems are designed to follow the "query - response" or shortly look-up concept. Users with an information need enter queries into search systems and those search systems produce ranked lists of search results. Ideally those search results are relevant for the queries used [115]. Look-up tasks are among the most basic types of search tasks, usually happening in the context of question answering and fact finding, e.g. wanting to know when the famous composer Mozart was born or who is the inventor of Penicillin.

Users often face the situation where they cannot find an answer to their information need, and therefore, they have to browse, collect and review large amounts of documents and/or synthesize results from different sources [115] as also illustrated in the journalist example above. Another example is the following simulated work task that was also used during our experiments: "Find the best universities for your child wanting to study either architecture or political science, assuming you live in Germany and are able to support your child with 1500 EUR per month". I make the assumption that there is not yet a search engine available, which just takes the social profile of the parent and the child and their preferences and monetary possibilities as input and generates the corresponding result for this case. Hence, a single query-based approach

is usually insufficient for the tasks in focus in this thesis: *complex search tasks* (to be defined in Chapter 2).

Matching complex information needs to queries can be a cumbersome activity. Nowadays, a search for such a task will still involve a lot of *aggregation* of different information sources, *discovering* lots of new facts of what is important in such a search (like understanding that accommodation is a very important cost factor to account for apart from tuition in our university example), and the need to *synthesize* all this information in a manner allowing to make an informed decision [102]. The quest for more efficient information search tools is more relevant than ever and this dissertation is just one additional building block in understanding how to proceed.

The goal of this thesis is to improve the special kind of Web search that results out of complex information needs. We have developed a theoretical model for complex search processes. Based on this model we have developed tools to carry out user studies to evaluate the complex search process. The results of these evaluations are the basis for the suggestions to improve the support for complex search in Web search engines.

The topic of this thesis is broad and therefore the concepts and methods used originate from various scientific sub-disciplines such as information retrieval (IR), interactive information retrieval (IIR), question answering (QA), information seeking and also fields such as exploratory search, Web information retrieval support systems and usability. Hence assigning this work to a certain field or discipline is far from trivial. Overall the work investigates the usability of search engines for complex search tasks. It therefore uses many concepts and definitions from IR. To cover interactive aspects in our experiments and also make them measurable with our logging technology, we used concepts and methods originating from IIR. As far as the character of the search tasks in focus (complex tasks) is concerned, they are based on an amended definition that is still quite similar to the one used in exploratory search and QA related research - but has the advantage of being better measurable. Finally, when it comes to the results of this thesis, they would most probably be complementing Web information retrieval support systems research.

This thesis is organized as follows: In the next chapter, I introduce complex search and its role in Web search along with the definitions of the core concepts used in this dissertation. After the prerequisites have been introduced, I present the problem statement along with the research approach used in this dissertation. After the research approach I present the core research questions that guided my research. In the following section I answer the research questions, outlining the research and summarizing the contributing studies. Then I present the ATMS model (defined in Chapter 5 on page 67) to improve the

support for complex search and discuss the model along with conclusions and limitations. Finally I demonstrate the main steps taken during the development and implementation of the Search-Logger study framework in the appendix.

This is a "thesis by publications". The thesis is composed of an overview of around 100 pages, to which 7 relevant publications are attached. Those publications are a joint works together with my co-authors. I have added one paragraph to each publication summary to point out my own contribution.

Chapter 2

Definitions and nomenclature

Due to the ambiguity of some terms, a precise definition of the concepts used is necessary. As there are several definitions for many of the following concepts, I have in those cases selected one that is appropriate to be consistently used throughout my thesis.

Web Search Query

I have found the following definition from the IT Law Wiki [2] well suitable for my needs in this dissertation:

> "A *web search query* (also called a *search query*) is the string of characters that a user enters into a search engine to satisfy his or her information needs."

Search Session

I use the definition by Jansen et al. [51, p. 862]:

> A *search session* is "a series of interactions by the user toward addressing a single information need".

The session typically starts with the first interaction of the user with the search system and ends, when the user either (1) successfully, (2) partly successfully or (3) unsuccessfully (has given up) leaves the system or starts another session.

Task

The following definition is based on Ingwersen and Järvelin's book "The Turn" [46] and a paper by Li and Belkin [70]:

> A *task* is an abstract description of activities to achieve a certain goal.

The concept of a task in the context of search (evaluation) can be decomposed into a work task and a search task resulting out of this work task [11, 45, 12].

Work Task

The definition best suitable for the needs in my dissertation can be found in Ingwersen and Järvelin "The Turn" [46, p. 392]

> A *work task* is "A job-related task or non-job associated daily-life task or interest to be fulfilled by cognitive actor(s). Work tasks can be natural, real life tasks or be assigned as simulated work task situations or assigned requests."

Simulated Work Task

Also the following definition is from Ingwersen and Järvelin "The Turn" [46, p. 390]:

> *Simulated work tasks* are "Work task/Interest situations designed for IS&R (information seeking and information retrieval) research by involving a specified but artificial scenario or cover story of semantic openness. The situation at hand is meant to trigger individual information needs in test persons in a controlled manner."

Search task

For this definition I also refer to Ingwersen and Järvelin's book "The Turn" [46, p. 73]:

> "A *search task* is a sequence of activities with the goal of finding specified information - the specification may range from narrow and detailed, e.g., a fact, to broad and vague, e.g. something about memory problems in old age."

This definition is specified further in the paper by Li and Belkin [70]:

> The search is usually carried out with IR systems.
>
> A search task usually comes out of a work task.

A family might for example be dealing with the task to plan a holiday trip. Resulting out of this work task might arise the search task to find children friendly hotels at a certain destination. According to Bell and Ruthven [12] the search tasks can either come from the searchers themselves or can be artificially created within laboratory evaluations.

Look-up/Simple Tasks

Look-up tasks are search tasks that lead to look-up searches. In this dissertation, look-up tasks and simple search tasks are regarded as being synonymous and defined by White and Roth in "Exploratory Search: Beyond the Query-Response Paradigm" [117, p. 13] as follows:

> "Lookup searches generally involve the retrieval of single answers (e.g., a single piece of information satisfies a known item search, fact retrieval, or question answering; a single Web page satisfies a navigational query submitted to a Web search engine)."

Look-up searches are among the most basic types of search tasks. Usually they happen in context of question answering and fact finding. Typically they are needed to answer who, when and where questions [117].

Complex Search Tasks

For this definition I refer to our own papers [98, 100, 101]:

> *Complex search tasks* are tasks where users are required to follow a multi-step and time consuming process that is not answerable with one query, requiring synthesized information from more than one retrieved web page or document to be solved. The process to work off complex search tasks usually comprises at least one of the process steps aggregation, discovery and synthesis.

To get a better understanding, an example for a complex search task was described and explained in Chapter 1 on page 12.

Complex Search Behavior

Complex search behavior appears when people carry out complex search tasks. The actual execution or execution process of the search task, independent of fulfilling the goal to satisfy the stated information need or not, will be referred to as a search[1].

[1] This definition is analogous for the terms "simple search" and "simple search tasks"

Relationship between complex search and exploratory search

The goal of this thesis is to make time consuming search tasks that require a lot of user interaction measurable and to also investigate the labor-intensive aspects of the search process such as querying, tabbing, copying, and pasting. At the very beginning of our research we tried to use Marchionini's definition of *exploratory search* [79, 117]. Exploratory search tasks (see also Figure 2.1) are defined as open-ended, abstract and poorly defined information needs with a multifaceted character. They are usually accompanied by ambiguity, discovery and uncertainty. Such exploratory search tasks fulfill needs like learning, investigating or decision making and require a high amount of interaction [117].

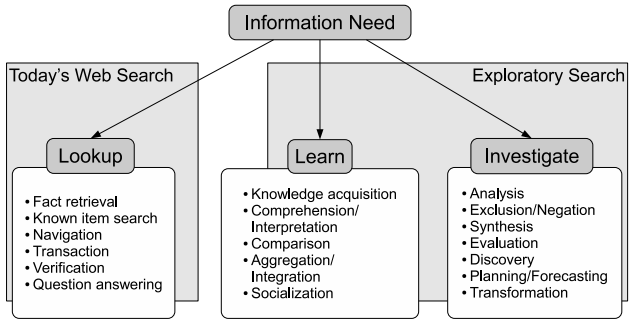

Figure 2.1: Types of search activities according to Marchionini [79]

Marchionini's exploratory search concept was a very good start and our chronologically first publication with the title "Search-Logger - Analyzing Exploratory Search Tasks" [102] still carried "exploratory search" in the title. In the course of the subsequent research we understood that this definition was not entirely suitable as it was based on too many cognitive concepts and overlapping activities.

In his view of non-lookup search Marchionini summarizes all sorts of search related activities under "learning" and "investigating". At closer hindsight those activities are located at different cognitive levels of the search process. While some concepts such as aggregation, discovery and synthesis are first level search and information gathering steps, activities such as planning/forecasting, analysis, or evaluation are of a different quality. They often use information that has been found in the first step, e.g. through aggregation and synthesis, and process that information to achieve the needed result. This makes such information processing steps second level search activities.

Some of the activities such as aggregation certainly are interactive, but the

amount of learning during pure aggregation tasks can be limited. Therefore aggregation and comprehension or knowledge acquisition, which Marchionini all put into a single group called "learning", are in fact quite different concepts as far as learning is concerned.

In addition quite a few of the activities overlap. For example analysis, comparison and evaluation are different concepts, but they also have a lot in common when being used to carry out a search task. For example a comparison task or an evaluation task will almost certainly contain analysis elements. Planning/forecasting is a high level, cognitive process that can be seen as being based on activities such as aggregation, evaluation and analysis.

Finally, one of the goals of my dissertation was that the final outcome of the thesis needs to be technically implementable. Core activities of exploratory search such as knowledge acquisition, planning/forecasting or interpretation are certainly important aspects of learning and investigating. Yet they are very broad and could therefore only be implemented for certain domains, but not on a general level.

Having understood the assets and drawbacks of Marchionini's exploratory search definition, I needed a different search model that does not contain the second level search activities such as knowledge acquisition, comprehension or planning/forecasting as I would not be able to measure those. I was therefore looking for a model of interactive, labor-intensive, time consuming search that is measurable and can be implemented. We used Marchionini's exploratory search model and carefully investigated the activities he mentioned. We set the scope on the first level search activities and left out any second level information processing activities as we expected them as being too difficult to be implemented on a general, non-domain specific level. We found a model based on aggregation, discovery and synthesis as developed in [98] and further defined in [100, 101] as suitable for our needs. It covers the main interactive and labor-intensive steps of finding documents to known problem aspects, discovering new problem aspects and finally summarizing (in the simplest way copying and pasting) the found information such as URLS into one document. All activities are first level information gathering activities. The term that seemed to be most appropriate to describe that kind of search was "complex", as illustrated in Figure 2.2. Needless to say that the line between complex search and exploratory search as Marchionini defined it can be a thin one. Especially in cases when searchers make a lot of discoveries during searching. Complex search tasks require a lot of interaction with the system. Hence the search effort in terms of manual labor will be significant, but complex search tasks do not implicitly carry all the standard attributes of exploratory search tasks such as learning, planning and decision making [115]. They are therefore complex,

but not necessarily exploratory as also outlined in Figure 2.2. I will give an example further below.

In addition our Search-Logger is designed to measure user action that is directly related to the labor intensive aspects of search processes to fulfill complex information needs such as time effort, number of queries, pages visited and browser tabs opened and closed. The Search-Logger currently does not allow to measure parameters such as engagement and enjoyment, information novelty, or learning and cognition that were mentioned by White [116] as being appropriate measures for exploratory search tasks. Hence it would not be correct to say that the Search-Logger is capable of measuring exploratory search tasks if its measuring capability really only covers certain activities of the exploratory search definition.

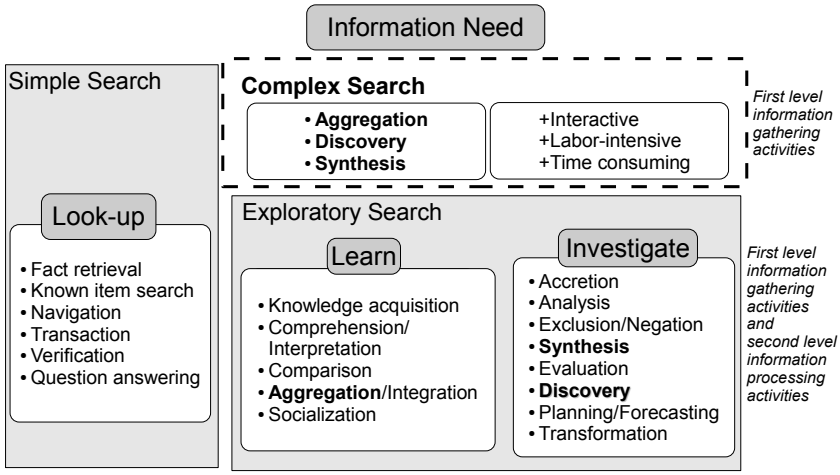

Figure 2.2: Comparison of complex search vs. exploratory search (based on [79] and extended)

The difference between our understanding of complex search and exploratory search is best illustrated with the help of the following example task. Assume you would be planning a holiday trip to the Canary Islands. As you have children, you want to find child-friendly hotels. You are using search engines to find those hotels. This task will take you a while. It will mean a lot querying and gathering links to hotel Web sites and also copying and pasting relevant hotel specific information, along with prices into a separate document. You might even observe that you could use the Spanish term for child-friendly, which is "apto para niños", for your queries. This would be a search consisting of aggregation, synthesis and even some discovery element. All those are first level search and information gathering activities. This process will certainly

be interactive and labor-intensive. But it does not necessarily require a lot of learning and investigating. It starts becoming exploratory when you begin acquiring knowledge that e.g. Majorca, part of the Balearic Islands, might be a better choice for you traveling with children - as the child-friendly hotel sector is developed better there. You would then look for child-friendly hotels in Majorca. You would compare all the results that you have gathered, analyze and evaluate the offerings, exclude certain offerings and in the process learn which offering might be best suitable and make a decision based on the knowledge you have gained. This step adds the second level information processing activities to the search process. By means of this example it also becomes clear that complex search can lead to exploratory search if second level information processing activities such as decision making are also used in addition to first level information gathering activities.

Chapter 3

Research approach

3.1 Problem statement

Search engine quality measurement initiatives have widely contributed to enhancing search engine performance for general-purpose, fact-based queries. However the same is not yet true for all other contexts or information needs like complex search tasks [115] as defined earlier in Chapter 2 on page 17. Similar to the earlier mentioned supplementation of classic online search with complex search in order to cover all of today's search needs, the methodologies to evaluate the corresponding tools and services also need to be extended. Classic evaluation methodologies predominantly focus on the search system itself, not on the search process that people follow [67]. The evaluation is still limited to those systems that rely on minimal human-machine interaction [60]. In addition, user related aspects such as search proficiency, become important to determine search task success [74, 93] and have to be taken into account along with integrating the behavior of the user into the evaluation process of search systems [115].

According to a keynote speech with the title "Search isn't Search" by Stefan Weitz, Microsoft, given at the SMX 2009 Conference [113], only 1 in 4 queries is successful and many queries yield terrible satisfaction (based on data from Microsoft's Bing search engine). Many search "queries" are actually tasks (as defined in Chapter 2 on page 15), close to 50% of the tasks are longer than 1 week and people are increasingly using search to make decisions (66% of search users). As a result, today's Web search engines are only being appropriately used for a subset of tasks that they could theoretically be used for.

Search engines like Google or Bing have started adding more features to their Web search systems to better support complex information needs, such as automatic query extension or universal search. The universal search model [106],

first revealed in 2007, integrates document surrogates, videos and images in the first search result pages and therefore offers a more complete and comprehensive user experience. Yet these features still add little incremental support [98] for carrying out complex search tasks as defined in Chapter 2 on page 15.

The research presented in this thesis is aimed at understanding complex search and improve its support in current Web search engines. To guide the research process, I have divided the dissertation into the following sub tasks and corresponding research questions:

3.2 Research steps

The problems and challenges mentioned in the above section made a combined approach necessary. As complex search is not an established research discipline such as information retrieval or interactive information retrieval, but a research topic that can be investigated in several disciplines, I needed to study those related research disciplines to build a common ground and develop a definition for complex search. Based on this definition and an according model for complex search I developed a method to measure and analyze complex search.

I used this method to conduct user studies with the goal of describing complex search behavior with simple measures. Based on theses findings I made recommendations for improving the support for complex search in search engines. As outlined in Figure 3.1 the 4 steps in my research approach were:

Step 1 - Study established search models for non-simple search tasks and research approaches in related areas

Step 2 - Select a method and build the relevant tools to measure and analyze complex search tasks. Test applicability of the method and the tools in the course of a pilot study. If needed go back to previous step and revise method.

Step 3 - Carry out user studies and learn how users search and what impacts their search performance

Step 4 - Analyze the results of the studies and give recommendations for improvement of the support for complex search tasks in Web search engines

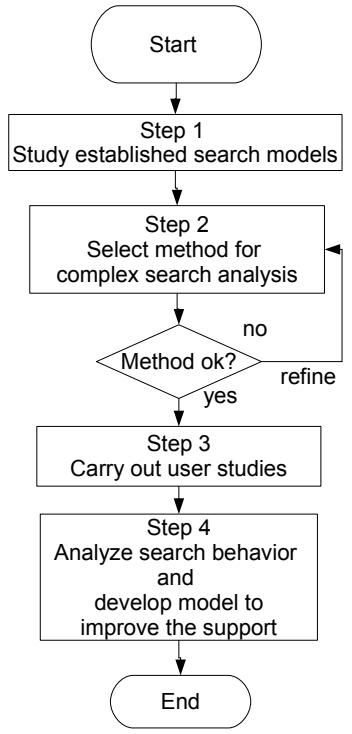

Figure 3.1: Flow chart illustrating my research approach

In the following sections I will describe the 4 steps in more detail.

Step 1 - Study established search models

To understand complex search and to develop a measurable model for it, I analyzed existing search models and I will present a clear break-down of complex search into measurable steps. The findings are published in Publication I [98]. We took the exploratory search model by Marchionini [79] as a basis and developed a reduced model (omitting the second level information processing activities such as planning/forecasting) for complex search based on the better measurable first level information gathering activities aggregation, discovery and synthesis.

When increasing the support for complex search, the whole notion of complex search needed to be made measurable. The approach that I followed in this dissertation was looking at existing research directions in related disciplines and selecting elements that can be used for measuring complex search. I screened traditional and established research approaches in information retrieval (IR),

interactive information retrieval (IIR), exploratory search and open domain question answering with a focus on ways to carry out user studies with complex search tasks and also on potential measures to describe complex search behavior.

IR research is focused on systems, which retrieve relevant documents from a document collection [6]. The focus of research is on the system and users are not part of the scope of classic IR. Researchers in IR examine how people retrieve information from repositories in commercial organizations, in public places like libraries but also on their PCs [115]. The query - answer concept (users enter a query into a search field and get back a ranked list of search results) used in IR has been the basis for commercial search engines like Google, Bing or Yahoo.

The study methods used in IR can mainly be divided into two groups: juror based studies and click-through data based studies. In juror based study methods to measure the retrieval effectiveness of IR systems, a static set of documents plus a set of queries is taken and according results are evaluated by jurors (see [65]). This so called Cranfield methodology [22] is widely used for evaluating IR systems. The Text Retrieval Conference (TREC) chose it as the main paradigm for its activities of large-scale evaluation of retrieval technologies [112]. Those methods evaluate search engines and their performance on a technical level, while leaving out important user related aspects. Researchers try to integrate the user into the measuring process as can be seen in TREC Interactive Track [27] or TREC High Accuracy Retrieval of Documents Track [3].

The two most commonly used measures in IR research are precision and recall plus many additional derived measures like normalized discounted cumulative gain, reciprocal rank or expected browsing utility (see [66]). While with precision measures the relevance of results displayed, recall is a measure for how well the set of retrieved documents covers the total set of available and relevant documents in the collection.

IIR focuses on the search process itself (and not the used system) and tries to especially overcome the artificial distinction between user and system in information retrieval evaluation efforts [91]. The hypothesis is that search is very often not only query-response but an interactive process. Ruthven [94] argues that most information seeking is usually part of a bigger search task (not only a session or a query) or even a work task as described by [46] and analyzing the interaction with the system cannot be done without also accounting for the underlying task. As a task can usually span longer time frames (from minutes to weeks [113]), only examining parts of it without considering the whole context would not provide the right insights for the researcher.

When it comes to evaluating the performance of IIR systems, according to Kelly

[56, p. 27] studies can either be conducted in a laboratory environment or in a naturalistic setting, e.g. by using real search engine logs. Both ways have their advantages and disadvantages. While naturalistic studies are supposed to show more realistic user behavior, the researcher usually has little influence on the setting and therefore the results are difficult to compare across sites [56]. Naturalistic studies have the advantage that they can span longer periods of time (longitudinal studies). Data is usually collected in different ways. Kelly [56] mentions think-aloud methods (study participants express their acting during the experiment), stimulated recall (gathers the same info as in previous method but after the task has been carried out), observation (researchers observe study participants), logging (user activity is automatically recorded), questionnaires (a set of closed and open questions administered to study participants) and interviews (mostly open questions).

IIR experiments are usually also defined and differentiated from other experiments by what tasks the users are expected to carry out and on what corpora (document collections) the tasks are to be carried out. On the corpora side Kelly [56] mentions test collections (like TREC [27] and HARD TREC [3]), the Web as an information body, or natural corpora (collected by users and mainly used for personal information management studies). On the task side, Kelly mentions natural tasks (based on information needs users have in their everyday life), tasks that allow multi-tasking studies (e.g. [24]) and simulated work tasks [14, 46].

The measures used in IIR experiments can mainly be grouped into measures of context, measures of interaction, measures of performance and measures of usability. The overall measure to be maximized is user satisfaction [19].

The term "exploratory search" (ES) was coined by Gary Marchionini [79]. He divided the search task universe into look-up, learning and investigating. Exploratory search comprises tasks that require learning and investigating.

As far as evaluating exploratory search systems is concerned, the methods used are not as well-researched yet as they are in IR [115]. Similar to IIR, the challenge is that the user and the system need to be simultaneously examined [115] (as opposed to IR where only the system is in focus). The consequence is that the experiments usually are costly. They are often conducted with insufficiently big sample sizes and small numbers of tasks. Often the user samples are not representative, e.g. consist only of students.

The main shortcoming so far is the repeatability and comparability of experiments. A way to improve the comparability of an experiment is to set up test collections of tasks (similar to the ones used in TREC). According to Kules and Capra [63, p. 419], such a task should be structured as follows: It "(1) indicates uncertainty, ambiguity in information need, and/or need for discovery;

(2) suggests knowledge acquisition, comparison, or discovery task; (3) provides a low level of specificity about the information necessary and how to find the required information; and (4) provides sufficient imaginative context in order for the test persons to be able to relate and apply the situation."

Researchers have suggested developing special measures for exploratory search systems experiments. White [116] mentions the following measures as appropriate: (1) engagement and enjoyment, (2) information novelty, (3) task success, (4) task time and (5) learning and cognition. In addition, time is an important aspect not to be neglected when trying to evaluate complex search systems. Exploratory search sessions can span over days and weeks and usually comprise various activities [115]. Long term studies are essential in order to get realistic study results [57].

Finally I have investigated how researchers evaluate systems in the field of open-domain Question Answering (QA). As opposed to IR systems, which retrieve whole documents, QA tries to find correct answers to questions within documents [34]. QA techniques can be used to answer simple fact based questions, but have also been applied to find answers to quite complex questions, where the answer needs to be constructed out of multiple documents [111]. QA systems are usually built of four main modules that are set up in a chain: the question analysis module, the module searching for documents and analyzing them, the module that selects the relevant passages and finally the module that extracts the answer [30]. The QA track of the Text REtrieval Conference (TREC) deals with measuring the performance of QA systems. In TREC-8 each participant got access to a document collection and a set of 200 fact-based questions [111]. An example for such a question is "How many calories are there in a Big Mac?" [111, p. 83]. The participants answered the questions with their systems and submitted a ranked list of 5 answers per question. The answers were assessed by humans on a binary scale. It was guaranteed that for each question at least one document in the collection contained the answer. In the TREC 2003 QA track, tasks to create lists and find definitions were added to the questions, such as "List the names of chewing gums" or "What is a golden parachute?" [25, p. 1]. In the TREC 2006 QA track complex, interactive question answering (ciQA) tasks were added [25]. Here the performance of QA systems was measured also considering interaction aspects. In general the performance of QA systems depends on the complexity of the task and on how difficult it is to extract the answers [83]. Such systems can usually answer fact-based questions quite successfully. Yet they did not perform that well when more advanced linguistic techniques are needed [111].

Step 2 - Select experimentation method

I opted for the following combined approach to analyze the characteristics of complex search: Automatic questionnaires and user experience sampling.

The design of this method to measure those open-ended search tasks was specifically triggered by our own academic setting. As researchers we are often faced with the need to find solutions for vaguely defined problems. The quest for appropriate solutions usually starts with searching the web for hints and proxy information. Search sessions often end with users being annoyed by current search means and their inability to reflect those open-ended search needs. I was looking for a logging-based method to measure search tasks, which often might consist of several search sessions. In addition I wanted to be able to turn the logging functionality on and off at any time. Finally I wanted the method to be affordable and easy to use over the Web - ideally with any computer or platform.

The chosen method combines automatic experience sampling (e.g. via browser plug-in [31]) and automatic questionnaires - methods usually used separately in IIR studies (as outlined earlier). According to Boyce et al. [15, p. 202] analyzing the search process and user characteristics independently is well suited for performance predictions or to analyzing performance differences. As the user studies will mainly be performance-based, this approach perfectly meets the requirements.

The user interaction data, which is automatically gathered, is used to analyze interaction and performance aspects and contextual and usability aspects are investigated by taking into account explicit user provided data from the questionnaires as outlined in Table 3.1.

User sampling for this approach needed to meet the usual requirements for usability studies like working with a reasonably sized sample of study participants, and caring for the right backgrounds of study participants. Empirical data from the experiments [102, 100, 101, 103, 99] shows that especially if tasks are complex, the data that is collected can show large degrees of variation. In studies aiming at showing differences between e.g. certain user groups or context variables like age or gender, it can be difficult to get enough signal to produce significantly different average mean values of certain users groups. To avoid this problem, the user sample either has to be chosen to be big enough, or a series of special studies have to be carried out with users of similar backgrounds.

The Web was used as the information body (instead of a specific document collection) as we were analyzing Web search behavior.

This method is specifically designed to carry out user studies with simulated

Experience sampling via browser plug-in	**Automatic questionnaires**
time based measures (task time, session time, search speed)	specific feedback on query, session and task level
interaction measures (number of queries, number of pages visited, query length)	demographic data (gender, age, education)
performance measures (bookmarks, data copied to clipboard, Web pages visited)	information need data (domain expertise, task difficulty and complexity)
	prior search experience (use of the Internet, use of search engines)

Table 3.1: Features of this combined approach

work tasks (for the definition see Chapter 2 on page 16) based on goals as described in [10, 14, 63]. Examples for such tasks can be obtained from Publication IV (p. 7) in the appendix of the thesis. Depending on the technical manifestation of the method (e.g. if an easily distributable and usable logging technology like a browser add-on is applied), the method can also be used to conduct studies with natural tasks as described by [56, p. 82], because users can easily install the logging technology themselves and no laboratory environment is required. Natural tasks are taken from people's usual common duties, like a researcher collecting references to write the related work section of a scientific paper. I developed a logging tool called "Search-Logger" to carry out user studies with complex search tasks according to the selected method (automatic questionnaires combined with user sampling via browser plug-in). We have published the tool in Publication II [102]. For technical details about the implementation of the tool please refer to Appendix A. I have identified several other tools designed mainly for the same purpose of logging user events like the "Wrapper" by Jansen et al. [48]. The Wrapper tool was developed for the purpose of logging events of all applications used by an information seeker (including applications like Microsoft Office or email clients). The concept of search task does not exist in this approach. Another tool is a browser plug-in that was created by Fox et al. in 2005 [31]. Fox's approach was implemented as an add-on for the Internet Explorer. It was the first tool to gather explicit as

well as implicit information during searches at the same time. It evaluates the query level and gathers explicit feedback after each query. Fox's approach also came with a sophisticated analysis environment for the logged data. Another tool was Lemur's Query Log Toolbar" [23] which is a toolkit implemented as a Firefox plug-in and Internet Explorer Plug-in. It logs implicit data on the query level. Logging complex search tasks is not implemented. Other similar tools are the HCI browser [16], The Curious Browser [21], WebTracker [108] and Weblogger [89].

While most of the tools were developed for evaluating the query level, the Search-Logger was especially developed for evaluating complex tasks. None of the tools have the built-in functionality to have a pre-compiled set of complex search tasks carried out by a test group in a non-laboratory environment without any time constraints.

Step 3 - Carry out user studies

Finally, to get insight into the characteristics of complex search tasks and the corresponding user behavior, I carried out user studies with different user groups together with colleagues. The first pilot study with student participants was conducted to test the Search-Logger tool. The Search-Logger tool along with the results of this pilot study were published in Publication II [102]. The results confirm that the Search-Logger tool is suitable to carry out user studies with complex search tasks - it correctly logs the right variables allowing me to analyze the appropriate measures. In addition it could clearly be shown that it takes users a considerable amount of interaction and time to carry out complex search tasks with current search engines. The next study was conducted with library search experts in the course of a library search contest. The findings of this study were published in Publication IV [104]. The most important observation was that all participants predominantly used search engine strategies (start the search process with entering queries into search engines) as opposed to using known address strategies (directly navigating to a known Web site other than a search engine). In addition the library search experts used a parallel-player strategy throughout the whole experiment, continuously having multiple browser tabs open and closing old ones and opening new ones - which is common for more complex search tasks. The third study was carried out with ordinary Web users (Figure 2 shows some images taken during this study). The findings of this study have been published in Publication III [101], Publication V [103], Publication VI [100] and Publication VII [99] and can be summarized as follows: The complexity of search tasks is measurable and its characteristics are significantly different from simple tasks. The complexity can be expressed by applying measures such as task time, session time and number of browser

tabs opened. A correlation exists between people's normal Web activities and their performance when carrying out complex search tasks with Web search engines. The more complex a search task becomes, the less people are able to judge certain task parameters such as difficulty, effort and outcome. Gender is an indicator for search performance when carrying out complex search tasks.

Figure 3.2: Pictures taken during the user study in Hamburg (myself, study participants, one of my supervisors Ulrich Norbisrath - from left to right)

Step 4 - Analyze data and give recommendations for improved support

I analyzed the data gathered during the user studies to develop a set of recommendations and best practices to improve the support for complex search with search engines. The outcome of this step is the ATMS (awareness-task-monitor-share) model to better support complex search with current Web search engines and increase user satisfaction (defined in Chapter 5).

In the next section I state the research questions that guided my research throughout my PhD project.

3.3 Research questions

Understand complex Web search

To understand why current search tools do not support complex Web search tasks as well as they do in the case of look-up tasks, it is important to clarify where the limits for their application are. This requires an in-depth analysis of the existing search paradigms, the ones used in information retrieval (IR) as well as the ones used in related disciplines such as exploratory search, interactive information retrieval (IIR) and question answering. For definitions of those research disciplines and concepts please refer to Section 3.2.

To find the limits of current search tools and to improve the support for complex search, I will address the following research questions:

RQ1.1 What are the tasks that can/cannot be carried out well with current Web search tools?

RQ1.2 Based on the concept of exploratory search [79] (and its partly difficult to measure aspects like learning, planning, and decision making), is it possible to develop a model for complex search based on measurable activities?

RQ1.3 How does this model relate to the exploratory search model [79] and what search tasks would it cover?

Develop a study method and corresponding tools

After building a better measurable model for complex search (as motivated in RQ1.2), the next step is to develop a method and according measures to make this model measurable. This step raises the following main research questions:

RQ2.1 Can a method to analyze complex search processes be developed and how should such a method look like?

RQ2.2 What measures should be analyzed to characterize complex search tasks?

RQ2.3 What tools are needed to conduct user studies using the previously developed method?

Carry out users studies

The next step is deploying the method and the corresponding tools developed in step 3.3 and analyzing what complex search behavior looks like and what characteristics can be identified. This step raises the following research questions:

RQ3.1 Using the measures established in RQ2.2, what distinguishes complex search tasks from simple search tasks?

RQ3.2 How is successful search behavior reflected in those measures?

RQ3.3 Are there best practices how users carry out complex search tasks with Web search engines?

RQ3.4 Is there a relation between people's internet habits and their performance when carrying out complex search tasks?

RQ3.5 How do users perform when assessing difficulty, effort and outcome for carrying out complex search tasks with current web search engines?

RQ3.6 How does gender impact complex search performance?

RQ3.7 How does age impact complex search performance?

Improve support

Finally, after having analyzed the complex search behavior, the results are used to give recommendations for improving the support for carrying out complex search tasks with Web search engines, raising the following main research questions:

RQ4.1 How can the support for carrying out complex search tasks with Web search engines be improved?

RQ4.2 What are the general limitations of the support that can be given? What service level makes sense for what group of users?

Vision

My dissertation can only be an incremental building block in the research towards a wider vision of a "decision support and learning engine" that will "find" answers to problems and supports decision making and learning, liberating the user from the cumbersome searching and information collection tasks. Such a system would present all the necessary data and information in way that allows the user to make a decision based on all the relevant information (theoretically) available.

Thesis outline

The rest of the thesis is organized as follows, taking the seven contributing studies into account:

- Chapter 4 summarizes the seven publications included in this thesis
- Chapter 5 presents the ATMS model to improve the support for complex Web search
- Chapter 6 states the conclusions and limitations
- Appendix A contains a summary of the technical implementation details of the Search-Logger tools
- Appendix B contains the original publications as they were published or submitted for review

Chapter 4

Summary of publications and contributions

This PhD thesis refers to the following 7 scientific papers by myself and my co-authors that have been published in journals and conference proceedings (or are currently in press or under review):

- Publication I: "Complex search: Aggregation, Discovery, and Synthesis" [98]
  - Presents the literature review analyzing established models to describe interactive search processes
  - Suggests a search model based on the concepts of aggregation, discovery and synthesis
- Publication II: "Search-Logger - Analyzing Exploratory Search Tasks" [102]
  - Presents a tool to evaluate the user behavior when carrying out complex search tasks
  - Presents the results of a pilot study
- Publication III: "Ordinary Search Engine Users Carrying Out Complex Search Tasks" [101] (submitted for review)
  - Proves that complex search has special characteristics that are measurable
  - Presents measures that make complex search tasks distinct from simple tasks

  - Presents the different characteristics of correctly and wrongly carried out search tasks and good and bad searchers

- Publication IV: "Search Strategies of Library Search Experts" [104]

  - Presents search strategies deployed by library search experts
  - Presents an analysis of impact of Internet user types on search performance
  - Presents an updated classification of Web search strategies

- Publication V: "The Relationship between Internet User Type and User Performance when Carrying Out Simple vs. Complex Search Tasks" [103]

  - Presents the correlation between Internet user type and Web search performance for simple tasks
  - Presents the correlation between Internet user type and Web search performance for complex tasks
  - Presents the user-type-specific difference in performance between simple and complex tasks

- Publication VI: "Ordinary Search Engine Users assessing Difficulty, Effort, and Outcome for Simple and Complex Search Tasks" [100]

  - Presents how ordinary web users perform when assessing difficulty, effort and outcome for carrying out complex search tasks with Web search engines

- Publication VII: "Impact of Gender and Age on Performing Search Tasks Online" [99]

  - Presents gender and age differences when carrying out simple and complex search tasks.

In the following section I will briefly summarize each publication by stating the research method, the results, the related work and limitations and future work. As I have written those papers with co-authors, each paper summary contains one subsection about my own contributions.

4.1 Publication I: Complex Search: Aggregation, Discovery, and Synthesis

This paper [98] analyzes established models in exploratory search, information retrieval and interactive information retrieval and suggests a model for "complex search" based on the relatively clearly defined and measurable concepts of aggregation, discovery and synthesis. Aggregation can for example be measured by the number of documents found for a certain aspect of a search need. An approach to measure discovery could be to count the number of aspects (or dimensions or facets) of a specific search need that an information seeker spots during a search. Synthesis could e.g. be made measurable by comparing the size (in terms of number of words) of a set of relevant documents with the size of a summary report of those documents.

The paper was mainly motivated by the following questions:

- What are the main information gathering activities during complex search processes and can a model for complex search based on those activities be developed?
- How well is such a model for complex search supported by the current Web search tools?
- How can this model be mapped to existing search models?

The study answers research questions RQ1.1 (What are the tasks that can/cannot be carried out well with current web search tools?), RQ1.2 (Based on the exploratory search concept [79] is it possible to develop a model for complex search based on better measurable activities?) and RQ1.3 (How would this model relate to the exploratory search model [79]?). The main scientific contributions of the publication are:

- 3-step process model for complex search
- Mapping between this complex search model and the exploratory search model by Marchionini [79]
- Analysis of the support of current Web search engines for this complex search model

My contribution

As the first author of this paper I contributed major parts to the model development, being responsible for the literature review and leading the process that

resulted in identifying aggregation, discovery and synthesis as the main (search process and information gathering related) steps when users carry out complex search tasks. I was the driving force in discussing the problems encountered with Marchionini's definition. I also carried out the analysis to what extent current search tools support aggregation, discovery and synthesis. I over-viewed the paper writing process and I contributed major parts of the related work, almost the entire Section 2 and considerable amounts of Sections 3 and 4. I added all the changes required during the final two revision cycles.

Research method

For this paper we carried out an extensive literature review, followed by descriptive case studies about how the three process steps aggregation, discovery and synthesis are supported by current search tools.

Results

Current search tools support simple fact-based, look-up tasks well. Their support is less good in the case of more complex information needs. We present a "complex search" model based on the time consuming activities of aggregation, discovery and synthesis. The model mainly says that any complex search task can be decomposed into an aggregation step, discovery step and synthesis step. Each of those steps is relatively well measurable (as shown in [98]) and therefore this model is the answer to research question RQ1.2 (see above).

Main findings:

- Aggregation, discovery, and synthesis, are the main first level information gathering activities in the complex search process
- The aggregation step of complex search is supported to some extent by many current systems; support for aggregation is just in its roots and improving it would serve users better
- The discovery step is not supported in standard web search interfaces but some support is given by advanced search tools
- The synthesis step is not supported in present mainstream Web search systems

Table 4.1 summarizes our findings regarding the support of aggregation, discovery and synthesis in current search tools.

	Aggregation	Discovery	Synthesis
Standard web search interfaces	yes	no	no
Dynamic query interfaces	yes	yes	no
Faceted browsing	yes	yes	no
Collaborative search tools	yes	yes	no
Social search	yes	no	no
Universal search interfaces	yes	yes	no

Table 4.1: Support of aggregation, discovery and synthesis in current search systems

Related work

Aggregation, discovery and synthesis have also been studied by other scholars. Aggregation of information in the context of search can be described by activities like selecting, storing, and maintaining data objects. Heyman et al. or Krause et al. [41, 61] have studied social bookmarking services like del.icio.us, which provide users with already aggregated collections relevant to some certain topic. These collections can become a great help if the topic corresponds to the search task of the search session [118].

Information discovery is referred to as generating new ideas while browsing relevant information [58]. Information discovery is one of the key elements of information search, on and off the web. A very specific aspect of discovering information relates to information seekers entering a domain new to them. There are different approaches to support the discovery of formerly unknown aspects, some are automatic approaches others are human-supported ones. El-Arini and Guestrin [29] offer an automated approach for discovering additional aspects to a specific information seeking problem. The information seeker can take a set of relevant papers and use this set as a "query". The system will then suggest additional papers that "the information seeker should also read". Koh et al. have developed a system to support creativity in learning by enabling information discovery and exploratory search [59]. With their system called "combinFormation" users are able to make clippings of found documents, arrange them on the screen, and also add conceptual relationships.

Synthesis is commonly referred to as combining parts of separate items into one single and new entity. Nenkova et al. [86] state three main tasks in a summarization process, which are (1) content selection, (2) information ordering and (3) automatic editing, information fusion and compression [7]. Regarding content selection, a key challenge is to find the topic that a document is about. In order to find the topic of a content, researchers mainly apply mathematical

models, amongst others word frequency models [76, 87], lexical chains [8, 96], latent semantic analysis [105, 37, 33] and content models [38].

Limitations and future work

We assume that improving the support for the activities of aggregation, discovery and synthesis during complex search tasks will tremendously help both the inexperienced but also the experienced user. In our three step search model based on aggregation, discovery and synthesis we deliberately omitted aspects such as planning/forecasting or comprehension/interpretation as those concepts are difficult to measure (compare also [117]). In the future it would make sense to integrate more of Marchionini's exploratory search activities [79], such as analysis, exclusion/negation or comparison into the model.

4.2 Publication II: Search-Logger – Analyzing Exploratory Search Tasks

This paper [102] presents a tool to evaluate the user behavior when carrying out complex search tasks. The Search-Logger is an experimentation method especially designed to carry out search task based experiments. The Search-Logger's architecture is designed around the earlier given definition of a search task (see Chapter 2). The Search-Logger implementation answers research question RQ2.1 (Can a method to analyze complex search processes be developed and how should such a method look like?), RQ2.2 (What measures should be analyzed to characterize complex search tasks?) and RQ2.3 (What tools are needed to conduct user studies using the previously developed method?).

The main scientific contribution is:

- Development of the Search-Logger tool that allows complex search user studies on the task level.

My contribution

I was the first author of this paper. I was leading the extensive development process for the Search-Logger tool and planned and carried out the pilot user study to validate the tool. This means, I acquired the study participants, I developed and collected the work tasks, guided the study participants in setting up the Search-Logger framework at their computers, distributed the tasks and I also analyzed the data gathered during the pilot study. Apart from the long and challenging development of the Search-Logger tool especially gathering a set of suitable work tasks was tricky and needed a lot of testing and fine tuning. I wrote major parts of the paper. I first submitted it to the A-rated conference CIKM 2010. After it was rejected I developed a major revision of the paper, taking into account all reviewer comments and resubmitted it to SAC where it got accepted.

Research method

I carried out a literature review on tools to measure complex search tasks. Based on this literature review we built the Search-Logger tool and carried out a pilot user study to test the tool.

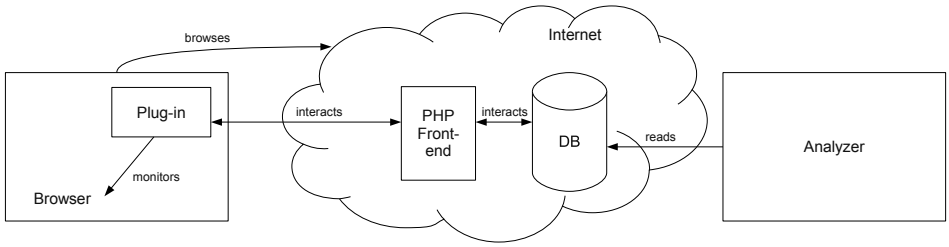

Figure 4.1: Search-Logger architecture

Technical details

The Search-Logger is realized as a browser plug-in (developed in Java-Script) for Firefox, completed by a remote log storage database and an analysis environment as outlined in Figure 4.1. It fulfills the following three main tasks: (i) administers pre-compiled search tasks to users, (ii) gathers implicit information about the search process by logging various browser events as outlined in the next paragraph, (iii) gathers explicit user feedback via standardized questionnaires supplied before and after each search task. With this approach we manage to log the search process on the search task level. Each logged event is tagged with task specific information like task name and task number and a time stamp. Based on this information the task performance can be analyzed and evaluated.

All data is centrally collected at a dedicated server. We can log the search process by gathering data on all measurable standard user events like total search time, number of web pages visited in total, number of browser tabs opened, search queries entered and number of search sessions started and ended.

To quantitatively analyze the log files, I used Excel macro programming methods as well as a few Python scripts (for more details on the analysis part please refer to Section A.3 on page 103)

Advantages of this tool overs existing tools:

- Task structure is implemented
- Tasks can interactively be administered
- Works across various platforms (Windows, Linux, MacOS)
- Easy to deploy and install
- Experiments can easily be set up and changed
- Low cost

Related work

I have identified several other tools designed mainly for the same purpose of logging user events. The first tool is called "Wrapper" by Jansen et al. [48]. The Wrapper tool was developed for the purpose of logging events of all applications used by an information seeker (including applications like Microsoft Office or email clients). The concept of a search task does not exist in this approach. Another tool is a browser plug-in that was created by Fox et al. in 2005 [31]. Fox's approach was implemented as an add-on for the Internet Explorer. It was the first tool to gather explicit as well as implicit information during searches at the same time. It evaluates the query level and gathers explicit feedback after each query. Fox's approach also came with a sophisticated analysis environment for the logged data. A third tool is Lemur's Query Log Toolbar [23] which is a toolkit implemented as a Firefox plug-in and Internet Explorer Plug-in. It logs implicit data on the query level. Logging exploratory search tasks is not implemented. Other tools are, the HCI browser [16], The Curious Browser [21], WebTracker [108] and Weblogger [89]. Most of the tools were developed for evaluating the query level while the Search-Logger was purely developed for evaluating longer lasting complex tasks. None of the tools have the built-in functionality to have a pre-compiled set of complex search tasks carried out by a test group in a non-laboratory environment without any time constraints and are therefore not suitable for our purposes.

Limitations and future work

At this point the Search-Logger only uses the browser add-on approach to record the user actions. Currently only a browser add-on for Firefox exists. This means that studies rely on Firefox browsers being used. In the middle term, the Search-Logger tool shall be extended to support more browsers but also different ways to log user activity. Apart from the existing browser add-on approach, a proxy-based approach is envisioned. Using a proxy (through which all user traffic during an experiment would be routed) would make it easier to set up experiments. Initial experiments with a proxy-based approach have shown that this approach also has its tricky parts like catching non-browser events or avoiding the logging of too many commercials.

In addition we are planning to merge the Search-Logger with the Relevance Assessment Tool by Lewandowski and Sünkler [68]. Adding the relevance dimension (on the search result level) to the Search-Logger will further improve the search engine analysis support.

We are also playing with the idea to run experiments with a mobile version of the Search-Logger (developed at the University of Tartu by a master student).

This mobile version extends a real browser with logging features on a mobile phone. Despite the high development effort, we currently do not see a real use case for complex search on mobile phones. Tablets might have an advantage to mobile phones in that regard, as they are bigger and more user-friendly.

4.3 Publication III: Ordinary Search Engine Users Carrying Out Complex Search Tasks (Manuscript submitted for publication)

This paper [101] presents the results of a study with 56 carefully selected ordinary Web users to investigate the characteristics of complex tasks. The study participants carried out a set of predefined simple and complex search tasks. The aim of the study was to examine the search behavior of ordinary Web users when carrying out those tasks and gain insights, which are valid for a widely valid part of the population in terms of age and gender. Many studies in this area are carried out with quite small user samples and users with backgrounds only from the academic sphere [72, 69, 55] (students and university staff). This raises doubts about the general validity of their results. We were especially interested in finding out more about (1) what makes complex search tasks distinct from simple tasks and if it is possible to find simple measures for describing complexity, (2) what are measures for successful search behavior when carrying out complex search tasks and if it is possible to distinguish good searchers from bad searchers by using selected measures.

We attempted to answer the following questions:

- What distinguishes a complex search task from a simple one?
- Can we identify successful patterns for search? What distinguished a successful searcher from an unsuccessful one?
- Can we make suggestions for search engine operators to improve the support for complex search tasks?

These questions and their answers are supposed to be one more building block towards better understanding the concept of complex search. At this point we have not derived them from a formal research evaluation but they are based on our personal observations. Along with the first results and the insight we got from them, we plan to iteratively put our future research questions on a more rigid basis.

This publication answers research questions RQ3.1 (Using the measures established in RQ2.2, what distinguishes complex search tasks from simple search tasks?), RQ3.2 (How is successful search behavior reflected in those measures?) and partly RQ4.1 (How can the support for carrying out complex search tasks with Web search engines be improved?). The main scientific contributions of this paper are:

- Proof of the significant differences between simple and complex search tasks and the presentation of the different measures
- Proof that successful and unsuccessful search behavior have similar characteristics and are hence difficult to distinguish with the measures we used
- Recommendations for improved support for complex search

My contribution

My contributions to publishing this paper as the first author were as follows: Prior to the experiment I was chiefly responsible for planning the experiment in terms of logistics (when, where, how) and raising the funds necessary to pay the study participants. While it is usually quite straight forward to get hardware purchases funded for IT research, it was especially difficult to get the financing for this user study in place. It took considerably longer than expected. I set up the Search-Logger infrastructure at the study premises in Hamburg and also over-saw collecting the set of simulated work tasks. During the experiment I was responsible for making sure that the experiment would run smoothly and without any unexpected incidents. After the experiment I developed the analyzer tool to analyze the data, I conducted the analysis of the data and I was chiefly responsible for getting the paper written. Especially the development of the analyzer tool was difficult and it took me a lot of iterations to produce the correct results. I was also responsible to coordinate the different authors and oversee the integration and revisions of the paper. I contributed almost the entire related work section, the results section, the method section and considerable parts of all other sections. I also carried out all revisions during the publication process.

Research method

We conducted a user study where we had ordinary Web users carry out complex and simple search tasks. We logged their search behavior and analyzed the data using standard quantitative techniques to find the relationship between independent variables such as task complexity and dependent variables such as search time or number of queries used by study participants.

Results

The complexity of search tasks is measurable and its characteristics are significantly different from simple tasks. The following set of measure has been

shown to allow characterizing complex search tasks and distinguishing them from their simpler counterparts:

task time: complex search tasks have a significantly higher search time; 427±17 sec. vs. 140±8 sec. in the experiment

number of sessions: complex search tasks are usually carried out in a higher number of sessions; 1.1±0.02 sessions vs. 1.0±0.01 sessions in the experiment

time on search engine results pages (SERPs): Users usually spend more time on SERPs during complex search tasks; 122±9 sec. vs. 33±3 sec. in the experiment

reading time: complex search tasks are characterized by users spending more time on reading and scanning pages; 307±15 sec. vs. 107±6 sec. in the experiment

number of pages visited: When carrying out complex search tasks, users usually visit a higher number of pages; 7.4±0.5 vs. 2.5±0.2 pages in the experiment

number of queries issued: When carrying out complex search tasks, users usually issue significantly more queries; 6.4±0.4 vs. 2.1±0.1 queries in the experiment

query length: The query length goes up with the complexity of the search task; 4.4±0.2 vs. 3.1±0.1 words in the experiment

number of query changes: The number of query reformulations correlates with the complexity of the search task; e.g. 2.0±0.1 vs. 1.2±0.03 new queries in the experiment

number of SERPs visited: Users visit a significantly higher number of SERPs when carrying out complex search tasks; 0.3±0.06 vs. 0.1±0.02 pages in the experiment

I grouped correctly carried out tasks (independent of which user carried out the task) and compared them with wrongly carried out tasks. The only measure that distinguished the correct tasks from the incorrect ones was the number of tabs - 2.9±0.2 vs. 2.4±0.1 tabs in the experiment.

Finally, I wanted to rule out the case where a searcher is excellent at one instance, but performs badly at all other tasks (which would not make him a good searcher and his search behavior would just randomly be good). The results show that good searchers can be distinguished from bad searchers by:

- smaller task times; 337±31 sec. vs 526±68 sec. for the best performing and worst performing quartile in the experiment
- smaller SERP times; 82±13 sec. vs. 166±29 sec. for the best performing and worst performing quartile in the experiment

Regarding the question of what makes a task complex, we can conclude that complexity can be expressed by the amount of effort (in terms of time, sessions, and queries) which is needed to carry out a search task. This can be shown and proven by means of the measures that we investigated.

We suggest that search engine operators put more emphasis on the fact that complex search tasks have significantly different characteristics than simple ones. These differences can be measured as shown in this paper and, depending on the character of the search task search engines could offer different kinds of support to the searcher. For example it would make sense to monitor the search process on the task level (as explained in Chapter 5 on page 67). When erratic or chaotic behavior from the user side is identified (like identical queries repeated), struggling searchers could be identified by giving hints or other forms of help. We suggest that a different, enhanced service should be offered to those struggling searchers (for more details better supporting complex search please refer to Chapter 5 on page 67).

Related work

I explored research aimed at investigating search user behavior. I found two main ways to analyze the user behavior: Using search engine transaction logs, and carrying out user studies in a laboratory setting. Logging user behavior was done in different ways. Transaction logs recorded in search engines have been analyzed in several publications [50, 80, 49, 39]. One of the first larger scale studies reflecting on the search behavior of Internet-Search users was published by Jansen et al. [52]. They analyzed transaction logs from the search engine Excite. In the paper itself it is mentioned that one of the disadvantages of that analysis is that the transaction logs contain no "information about the users themselves or about the results and uses" (p. 208). They talk about sessions, but it is not clear how their starts and ends are determined. As we have also done in this study, they analyzed the changes between queries. They distinguish unique, modified, or identical queries. Jansen and Spink [49] give an overview over nine transaction-log studies of five Web search engines based in the US and Europe. They review session length, query length, query complexity, and viewed content in different search engines. They observe that users are viewing very few result pages (even less then in their first study) and again that the use of Boolean operators is nearly insignificant.

Most of the presented studies are only transaction-based and do not take the actual task and user into account. Other studies (e.g. [18, 53]) were done with very specific user samples (like undergraduate students). All of them confirm that queries are generally very short (only around two words on average) and that users tend to only take a look at the very first results in the SERP. All studies giving tasks to the testing persons use simple tasks. The only exception is the study described in Hölscher and Strube [42], which investigated more complex search tasks. While most of the studies described use query logs from real search engines, some studies (e.g. [42]) log the behavior of certain users in a lab setting.

Limitations and future work

One of the limitations of this study was its sample size. Although our user sample of 56 people was much bigger than in most of the other studies we have found in the context of search studies, some measures were quite blurry - due to the diverse backgrounds of the study participants. We expect to see clearer signals and more significant differences between e.g. age groups or men and women in case of bigger sample sizes.

It would be interesting to further analyze the data concerning the possibility to identify patterns in the sequence of queries and if those patterns could be used to identify strategies and erratic search behavior. In addition it would make sense to more deeply investigate the differences between complex search tasks, where the complexity comes either from the effort to aggregate, or discover or synthesize. Finally we plan to carry out a similar experiment with a significantly larger user sample and more homogeneous user groups (like teachers or blue collar workers only) and investigate if it would be possible to get valid findings for the cases where hypotheses had to be rejected due to high standard error of the means. We assume that a larger user sample with more homogeneous backgrounds will lead to more significant differences.

4.4 Publication IV: Search strategies of Library search experts

In this paper [104] we present the results of a search experiment conducted with information professionals from libraries and museums in the course of a search contest. The aim of the experiment was to analyze the search strategies of experienced information workers carrying out search tasks of varying complexity.

We attempted to address the following questions:

- How do library search experts search?
- Can we identify unique strategies?
- Can we relate the Internet user types of the contest participants to their search performance?

This study answers research question RQ3.3 (Are there best practices how users carry out complex search tasks with Web search engines?) and partly answers research question RQ4.1 (How can the support for carrying out complex search tasks with Web search engines be improved?). Its main scientific contributions are:

- Presentation of search strategies of library search experts
- Analysis of the impact of the Internet user type on search performance
- Presentation of an updated classification of Web search strategies

My contribution

My personal contributions for this paper were planning the experiment from the technical and methodological side, providing simulated work tasks during the planning phase of the experiment, and carrying out the experiment together with colleagues - technically implementing and running the experiment. I considerably contributed content to the paper during the publishing phase. I drafted the abstract and introduction and contributed to the related work, method, results and conclusion sections.

Research method

We carried out a user study in the course of a library search contest (assigning the study participants work tasks that they had to carry out), logged the users' search behavior with the Search-Logger tool and partly automatically partly manually analyzed the data to identify common patterns in the search strategies applied by the study participants.

Results

The most important observation was that the participants predominantly (in 94.4% of the cases) used search engine strategies (start the search process with entering queries into search engines) as opposed to using known address strategies (directly navigating to a Web site other than a search engine). In only 5.6% of the cases, the library search professionals applied non-search engine strategies, navigating to known (non search engine related) Web sites and searching for the information there. This reconfirms that search engines are a good entry point to exploring a search space. The low average number of points (16 out of 30) that study participants reached during the experiment and the high average number of opened and closed tabs (49) and search attempts per task (62) indicate that users are required to search interactively and apply a considerable amount of manual labor to carry out complex search tasks of the kind used in the experiment.

Overall the library search experts most often applied the following strategies:

- Search engine strategy with subtype "search terms narrowing" (84.6%)
- Search engine strategy with subtype "search terms narrowing extending" (7.4%)
- Known address strategy with subtype "search terms narrowing" (3.1%)

Many of the contestants used a parallel-player strategy throughout the whole experiment, continuously having multiple browser tabs open and closing old ones and opening new ones, which is common for complex search tasks. It is interesting to observe that the results of the user study with 56 ordinary Web users [104] do not show this wide use of parallel-player strategies. The ordinary users worked with multiple browser tabs significantly less often - 4.9 tabs in the case of library search experts vs. 2.8 tabs on average in the case of ordinary users.

Regarding the Internet user type (a classification of Internet users into information seekers and communication and entertainment seekers derived from

people's online activities [54, 103]) it was interesting to observe that the participants who scored first, second, and third in the contest all have an Internet user profile "active versatile". This Internet user profile is the most active one, having high scores on all dimensions of Internet activity.

Related work

Search strategies have increasingly been researched in the last years. Marchionini [78] has defined four levels in information seeking: moves, tactics, strategies, and patterns. He defined strategies as generalized approaches to particular information seeking problems. Navarro-Prieto et al. [85] identified bottom-up, top-down, and mixed strategies. Chin and Fu [20] found in their study that younger users prefer the bottom-up interface-driven strategy. They look up more links and leave a Web page quickly. Older users prefer the top-down knowledge-driven strategy. Thatcher [107] has studied cognitive search strategies among experienced and less experienced web users. He identified 7 generic cognitive search strategies. Shneiderman [97] distinguished searching tasks from specific fact-finding to more unstructured open-ended general-purpose browsing tasks.

Kalmus et al. [54] define in their work the following types of Internet users: Active versatile (these are more active people, using different Internet possibilities like communication, information and entertainment), entertainment-oriented active (focus on searching for entertainment, and consumption of culture), practical work-related (focus on information and practical activities, active in using e-services), practical information-oriented small-scale (slightly higher than average use of information and e-services), entertainment and communication-oriented small-scale (searching for entertainment, communication, passive Internet use with regard to other purposes) and small-scale (not characterized by any specific Internet use, poorly developed online behavior).

Limitations and future work

The experiment was carried out with a time limit. Although we could not show a significant impact of the time constraint on the study outcome, the results might have turned out differently under open ended conditions as in the experiment described by Singer et al. in [102] where the study participants had 4 weeks to complete their tasks. We are planning an open ended follow up experiment with the same questions to further analyze the impact of the time constraint on the study results.

Overall the younger the participants were, the better they scored. We will also further analyze the correlation between age and search performance. As one

limitation of this study was, that the user sample only consisted of women, we will try to find a more balanced user sample for the next study.

4.5 Publication V: The relationship between Internet User Type and User Performance when Carrying Out Simple vs. Complex Search Tasks

In this paper [103] we present the correlation between people's Internet habits and their online search performance. It is widely known that people become better at an activity if they perform this activity long and often. Yet when examining Web search, the question remains whether being active in related areas like communicating online, writing blog articles or commenting on community forums correlate with a person's ability to perform Web search. Web search has become a key task conducted online. We present our findings on whether the Internet user type (defined below), which categorizes Web uses according to their online activities, has an impact on their search capabilities.

We attempted to answer the following questions:

- Is there a difference in search characteristics between certain user types when performing simple search tasks?
- Is there a difference in search characteristics between certain user types when performing complex search tasks?
- Is there a user-type-specific difference in performance between simple and complex search tasks?

This paper answers research question RQ3.4 (Is there a relation between people's internet habits and their performance when carrying out complex search tasks?). We show:

- Characteristics of different user types when carrying out simple search tasks
- Characteristics of different user types when carrying out complex search tasks
- Relation between Internet user type and search performance

My contribution

As the first author, I analyzed the data and compiled all the results for this paper. Especially the rigor of the statistical analyzes required, such as the Spearman's rho correlation coefficients, was challenging and deepened my knowledge

of statistics for user research. I over-viewed the paper writing process, I coordinated the participating authors and contributed major parts of the written content such as the entire results section and considerable parts of all other sections. I also conducted all revisions of the paper till it was accepted. For my specific contributions to the experiment itself please refer to the summary of Publication III, in Section 4.3 on page 46.

Research method

We conducted a user study where users had to carry out a number of work tasks and logged their search behavior with the Search-Logger tool. At the beginning of the experiment we also administered questionnaires to be able assign a unique Internet user type according to their usual Internet usage patterns. I used quantitative techniques to find correlations between the Internet user types (independent variable) and the search performance (dependent variable).

Results

To analyze the relation between Internet user type (for an overview of the Internet user type concept please refer the Publication V, included in this dissertation at page **??** ff.) and search performance, we have ranked the users according to their performance in the experiment - at first for simple search tasks only.

In the case of simple tasks, the average rank for Internet user type 1 (the most active one) was 2.7 versus an average ranking of 4.8 for Internet user type 5 (the least active one) on a scale from 1 to 10. Internet user type is an indicator for performance when carrying out simple search tasks. This finding was also confirmed by our correlation analysis. We investigated the correlation between the Internet user types and various search measures. The strongest correlation was between ranking number and Internet user type - also statistically significant. The higher the user type (the less active users are online) the lower they ranked in our experiment.

In the case of complex tasks the average rank for Internet user type 1 was 4.9 versus an average rank of 6.0 for Internet user type 5 on a scale from 1 to 10. We also investigated if there exist significant mutual differences between neighboring pairs (user type 1 vs. 2, 2 vs. 3, 3 vs. 4, and 4 vs. 5) but could not show any. However, we could show a significant difference of mean values when we grouped active user types (1, 2 and 3) in one group, and small-scale or more passive user types (4 and 5) in another (with a p value <0.03). This means that the active and more experienced Internet users perform significantly better

when conducting complex search tasks than the group consisting of small-scale Internet users.

Regarding a user-type specific difference in performance between simple and complex tasks, we observed that in the case of complex tasks, the average rank is significantly lower (indicating worse performance) for all user types in comparison to simple tasks. In the case of complex tasks the rank difference between Internet user type 1 and Internet user type 5 is 2.0, whereas in the case of simple tasks the difference is 2.1 - which shows that both less and more experienced users were equally struggling with the complex search tasks. I have run paired-sample t-tests comparing samples of complex vs. simple rankings for each user type. The resulting p-values (>0.5) indicate that the difference between the mean values of complex and simple rankings is not significant. As our user sample was comparably small ($n=60$) and the users had varying levels of search experience (ranging from inexperienced housewives to experienced students of information science), we assume that a larger sample in combination with a more homogeneous average search experience would lead to smaller standard errors and clearer results.

Related work

Different approaches are used to classify people according to their Internet usage behavior. For the purposes of this paper, we have conducted the Internet user classification based on the users' primary online activities. The original approach to this Internet user typology [88] examined a long list of different activities online, ran a factor analysis on those activities to distil the basic underlying patterns and later applied cluster analysis to determine the key types of Internet users. The following six basic user types could be confirmed:

1. Active versatile Internet users, who are active in both information seeking, and communication and entertainment related use;

2. Practical work oriented Internet users, who are mainly active information oriented users;

3. Entertainment oriented active Internet users, whose main interests include entertainment and communication related uses, and seeking Internet solutions that cater to their interests;

4. Practical information oriented small-scale Internet users differ from the previous group in so far as their activities focus mainly on information use, they are less frequent in their activities;

5. Entertainment and communication oriented small-scale Internet users also use the Internet less frequently than their active counterparts. Their use focuses on leisure-related activities and they are passive when it comes to information-related activities;

6. Small-scale Internet users use the technologies so infrequently that they do not have any significant types of activities that would describe them, and have poorly developed online behavior;

Future work

In future experiments, we are planning to use more nuanced user samples that allow us to compare two specific user types with each other and work out their differences. This comparison should provide the possibility to zoom in on the important activities. The analysis in this paper takes a statistical birds-eye view of the search process, while the data and data collection method enables us to investigate actual search patterns and search strategies. The browser based logging software would also enable us to follow the user in her or his natural search context and investigate the searches occurring naturally over the course of a day for a given user type. That data-rich natural experiment would give non-commercial tracking information and enable us to see search patterns carried across different websites and different periods of time. Supporting this with other methods (e.g. diary) would also enable us to look at cross-media search in an attempt to understand the searching of information in the context of other sources. The Search-Logger software would simplify this kind of cross-media approach as the search can be connected to other activities in the Internet browser.

4.6 Publication VI: Ordinary Search Engine Users assessing Difficulty, Effort, and Outcome for Simple and Complex Search Tasks

In this paper [100] we examine how ordinary Web search engine users manage to assess the four parameters difficulty, time effort, query effort and outcome for simple and complex search tasks. We compare according assessments for simple tasks and for complex tasks and also investigate, whether better searchers are also better judges. In addition we investigate, whether the judging performance depends on task complexity or simply on the individual searcher.

We tried to answer the following questions:

- How do users perform when assessing the effort for carrying out complex search tasks with current web search engines?
- Are there significant differences between assessing simple and complex search tasks?
- How does the users' ability to judge if the information they have found is correct or not depend on task complexity?
- Does the judging ability depend on task complexity or simply the individual user?

This publication answers research question RQ3.5 (How do users perform when assessing difficulty, effort and outcome for carrying out complex search tasks with current web search engines?). Its main contributions are:

- Users are good at judging simple tasks
- Users perform significantly worse when judging complex tasks
- Task complexity inversely correlates with judging performance
- Better searchers are not significantly better at assessing difficulty and effort
- Better searchers are significantly better at judging the task outcome for complex tasks

My contribution

For this this paper my contributions were as follows: I analyzed the data and computed all the results required for this paper. It was challenging to present the results in a way that they reflect the concepts difficulty, time effort, query effort and ability to find the right result in the most understandable and intuitive way. To make the text more understandable I finally decided to add a number of figures and tables to illustrate certain facts. As the first author I over-viewed the paper writing process, coordinated the contributing authors and almost entirely wrote the introduction, related work section, results section, and major parts of all other sections. I conducted all revisions till the paper was accepted for publication. For my specific contributions to the experiment itself please refer to the summary of Publication III, in Section 4.3 on page 46.

Research method

We carried out a user study where the study participants had to carry out a number of work tasks. Their search behavior was logged with the SearchLogger tool. Before and after each task we also asked the users to fill in task related questionnaires. Using standard quantitative techniques we analyzed the relationship between certain independent variables such as task complexity and various dependent variables such as judging performance.

Results

The results confirm that people are able to judge difficulty, time effort, query effort and task outcome for simple tasks. For 90% of the study participants the estimated and experienced difficulty was in line. 95% of the study participants correctly assessed their ability to find the correct result.

When examining users' ability to judge the aforementioned parameters for complex search tasks, as expected their ability decreased in comparison to simple tasks. 65% of the study participants were able to sufficiently judge the subjective difficulty. In addition, especially the high proportion of 73% of the users claiming to have found the correct results is not in line with our manual evaluation of their results. Only 47% (158 out of 336 carried out tasks) of the results that were submitted for complex search tasks were correct. This may indicate that the problem with complex Web searching might not be users finding no results, but the results found only seemingly being correct. This may explain why users are generally satisfied with their Web search outcomes.

When it comes to search capabilities, we would expect that better searchers are also better at judging the difficulty, the effort and the task outcome for complex search tasks. As the results show, only for the task outcome, users who perform better in the whole experiment were also significantly better at judging the outcome of the task. For difficulty and effort, the differences were insignificant.

Regarding the question whether the judging performance is independent of the task type (simple/complex) and just depends on the user, the answer is as follows: There were some users who were able to correctly judge the task parameters like task outcome (26% of all users) both for simple and for complex tasks. Yet the number of users who managed to correctly judge those parameters for simple tasks (and were wrong for complex tasks) was much bigger (51% in case of task outcome) than the number of users who correctly judged the parameters for complex tasks and at the same time were wrong with their judgments for simple tasks (only one user in case of task outcome). Although the numbers varied, this relationship also holds true for task complexity, time effort, and query effort. Together with the results from research questions RQ1 to RQ4 it is clear that task complexity impacts the judging performance of users.

Related work

Bell and Ruthven [12] carried out a user study with 30 people who were asked to work on three groups of search tasks (tasks organized in three complexity levels) and afterwards rate the complexity of each task on a 5-point scale. They observed that the assessment of completion and task complexity were inversely correlated. The more complex people perceived a task, the less confident they felt, when they completed that task. In addition they found that a task is perceived as more complex if the task contains little information about what information is needed and what amount of information should be retrieved. Also subjective factors like previous knowledge about topics related to that task had to be taken into account as influencing factors for the perception of complexity.

Gwizdka and Spence [36] conducted a study with 27 undergraduate psychology students in which they where required to fulfill a look-up task. They wanted to examine the relationship between searcher's activities and subjective post-task difficulty and finding predictors for subjective task complexity. They found that task time, time per click, pages visited, unique pages visited, revisit ratio and back-button use were good predictors for subjective task complexity.

White and Iivonen [114] conducted a study with 54 experienced Web searchers

and had them rate 16 search questions regarding complexity. Their results show that users perceive closed/predictable source questions easy, open/unpredictable source questions difficult. In addition the study participants agreed that "searchability, clarity, familiarity/currency, public knowledge, simplicity, and specificity" were important aspects that made a task either simple or complex.

Li et al. [71] conducted a survey containing 100 university students in China. They observed that objective task complexity measures were more indicative for task complexity than subjective ones. The main objective predictors for task complexity were: number of words in the task description, number of languages needed to interpret search results and the number of domain areas, that the task involved. In addition the objective complexity criteria were more helpful to predict complexity.

In the information science community the two concepts complexity and difficulty are sometimes used as being identical and sometimes they are used as being distinct. Gwizdka [35] has conducted a question-driven, Web based information search study with 48 participants (students, mean age 27 years) aimed at understanding the cognitive load when carrying out web search tasks (recording them and analyzing their respective actions). The study participants were required to carry out a primary task and in parallel a secondary task to measure their cognitive load on the primary task. His results confirm that subjective task difficulty and objective difficulty are in line and that study participants tended to underestimate task difficulty when being asked beforehand about it.

Vakkari and Huuskonen [110] conducted a study with 41 medical students to investigate how the search effort impacted search output and task outcome. They found that in case of bad retrieval results, their participants worked harder to achieve desired task outcomes. They conclude that measures for search process and task outcome need to be added to classic IR measures.

Future work

In future work it would be interesting to not only analyze if study participants correctly or incorrectly judged tasks but also investigate to what extent the users tend to over- and underestimate the task parameters. Regarding sample size we are planning to run experiments with bigger sample sizes. This will enable us to get more correct statistics with more significant features. In addition, we are planning to conduct studies with study participants from certain professional domains like teachers or blue collar workers only. Further on it would be interesting to investigate to what extent such abilities (e.g. task effort assessment) can be trained. The implication for a growing number of

information workers would certainly be important.

4.7 Publication VII: Impact of Gender and Age on Performing Search Tasks Online

In this paper [99] we examine gender and age differences for a user sample of 56 ordinary Web users carrying out a set of simple and complex search tasks. We compare the search performance of ordinary female and male search engine users when carrying out simple and complex search tasks and work out the differences. In addition we investigate the correlation between age and search performance for both simple and complex search tasks.

We tried to answer the following questions:

- What is the correlation between gender and search performance for simple tasks and complex tasks? What are the significant differences to simple tasks?
- Are women better searchers or men?
- How does age correlate with search performance for simple and complex search tasks? What are the main differences in performance between simple and complex tasks?

This publications answers research questions RQ3.6 (How does gender impact complex search performance?) and RQ3.7 (How does age impact complex search performance?). Its main contributions are:

- Women and men perform equally well for simple tasks (in terms of final results) and also show similar search behavior.
- In the case of complex tasks men and women also perform equally well, but their search behavior shows significant differences (in terms of SERP time, read time and browser tab usage).
- Younger people are quicker at searching than older people in case of simple tasks but both groups achieve similar results.
- In case of complex tasks, younger people are quicker and also achieve better results (higher number of correctly carried out tasks).

My contribution

I was the first author of this paper. I analyzed the gender - and age-specific data that we had collected during the study and computed the results needed for this paper. I over-viewed the paper writing process and almost entirely contributed

the written content. As there was very little related literature for gender and age specific aspects of Web search available, it was not easy to position the paper in the context of the existing work. I also carried out all revisions till the paper was accepted for publication. For my specific contributions to the experiment itself please refer to the summary of Publication III, in Section 4.3 on page 46.

Research method

We combined the information that we collected through automatic questionnaires with the automatically logged user behavior information gathered in the course of a user study. Using standard quantitative techniques we analyzed relationships between independent variables such as gender and age and dependent variables such as search performance.

Results

Our results show that men and women perform equally well when carrying out simple search tasks. The only significant difference between men and women carrying out simple search tasks was the number of pages visited. Men visited significantly more pages (2.8 for men vs. 2.2 for women). When it comes to complex search tasks, men and women again perform equally well. As opposed to simple tasks, the search behavior between the two groups showed significant differences. Men spent significantly more time on SERPs (145 sec. vs. 101 sec), issued a significantly higher number of queries per task (7.7 vs. 5.4), and women opened and closed a significantly higher number of browser tabs (3.8 vs. 2.6).

As far as age is concerned, younger (18-26 years age range) and older study participants (49-59 years age range) did not show a significantly different performance in case of simple tasks (both groups achieved similar quality search results). The significant difference between the two age groups was related to the time effort. Younger study participants had a significantly lower total task time (83 sec. vs. 186 sec.), SERP time (19 sec. vs. 40 sec.) and read time (64 sec. vs. 146 sec). For complex search tasks, also the search performance and hence the ranking was significantly better for younger users than for older ones (2.4 vs. 5.5). Also reading time (229 sec. vs. 434 sec.) and task time (333 sec. vs. 555 sec) were significantly smaller for the younger group than for the older group.

Related work

Lorigo et al. [75] used eye tracking to examine how different classes of users evaluate search engine results pages and found significant behavioral differences between men and women. Males checked more search results on SERPs in a more structured way. Jackson et al. [47] carried out a survey with 630 Anglo-American undergraduates to examine their Internet usage patterns and according gender differences. They found out that women were mainly using the Internet for communicating (e-mail), while men were mainly searching for information.

Hupfer and Detlor [44] carried out a survey-based study with 379 respondents, mainly students, to amongst others examine gender differences in Web information seeking. While women use the Internet for communication and are interested in finding medical information and information about government and politics, men seem to be more interested in hobby-related information and investment and purchasing information. Liu and Huang [73] did a survey with 203 completed copies at a University campus in China with people aged between 18 and 23. Their findings are that female readers prefer reading from paper to reading online, and that there are significant differences between what to read and sustained attention. Roy and Chi [92] conducted a study with 14 eighth grade students, 7 boys and 7 girls. The study participants had to carry out search tasks and were observed by two observers. Their findings show that boys used different search strategies than girls.

Meyer et al. [82] investigated the impact of age and training on Web search activity. In their study with 13 older and 7 younger users (ages not mentioned), they were able to show that the main difference between older study participants and younger ones was that both groups could fulfill most of the tasks, but it took the older ones more steps.

Morrell et al. [84] conducted a survey (consisting of 550 adults) to examine Web usage patterns among middle aged (aged 40-59), young-old (aged 60-74) and old-old adults (aged 75-92). They report distinct age and demographic differences in individuals who use the Web. Kubeck [62] examined the differences between older and younger adults finding information on the Web in a naturalistic setting. His sample consisted of 29 older (mean age of 70.6 years) and 30 younger (mean age of 21.8 years) people. He was able to show that both groups found answers of similar quality, but the older users where significantly less efficient in the process of searching. Aula [5] gave a set of search tasks to 10 older adults. She discovered that they were quite successful, but they had some operational difficulties in understanding how the Web was structured.

Dickinson et al. [26] present a prototype for a Web search system for older peo-

ple without any Internet experience. They also carried out a small user study and asked the users to rate the system against currently available mainstream search tools. The study confirms that older people search differently and have different requirements regarding user interface and usability.

Future work

One limitation was related to the broadness of our user sample. Due to the very diverse backgrounds (from university student to housewife), we were faced with quite high variances in our numbers. This resulted in high standard errors of mean. Hence we plan future experiments with more focused user samples (like a younger and older group of academics only) and those might produce more significant differences.

Chapter 5

The ATMS model for complex search with search engines

When setting out to improve the support for complex search, one should have in mind two main guidelines:

1. The user is in the center and the parameter that needs to be maximized is user satisfaction.
2. Improving any system specific aspects is meaningful as long as this improvement is positively recognized by the users.

According to Allan et al. [4] improving the system support for complex tasks is significantly more reasonable than trying to do the same for simple tasks. They have shown that for simple tasks users do not notice any improvements unless they are huge. This is in the contrary to difficult tasks with a lot of interactivity - even small improvements are perceived positively by users.

Overall, we have observed that task complexity can be expressed by the effort, which is needed to carry out a search task (in terms of time, sessions, browser tabs and queries). This can be shown and proven by means of the measures that we investigated. We found three main causes of complexity:

- It can originate from the need that a lot of information needs to be processed, read, qualified and collected.
- It can be due to the openness of a task where it is a priory not clear, what the main criteria of the task to be explored are, and a lot of effort is needed to discover the dimensionality of the result space and the main aspects of the task.

- It can result from the fact that the collected information needs to be synthesized into a single document and this causes a lot of effort.

5.1 Model outline

I have proved in this thesis that complex search is usually a tedious and time consuming process and complex search tasks are only weakly supported by current Web search engines. User satisfaction decreases the more effort users have to invest into a search task [113].

Hence minimizing a user's search effort by decreasing the search time, reducing the number of queries, the number of query reformulations and pages visited should improve user satisfaction. In addition, offering support along the process should theoretically also have a positive influence on user satisfaction. Based on the research carried out in this dissertation, the support that commercial search engines can offer for complex search tasks should be based on the following ATMS (Figure 5.1) model. ATMS is an acronym that stands for **A**wareness building, **T**ask features, **M**onitor search behavior and **S**hare best practices. The ATMS model also answers research questions RQ4.1 (How can the support for carrying out complex search tasks with Web search engines be improved?) and RQ4.2 (What are the limitations of the support?).

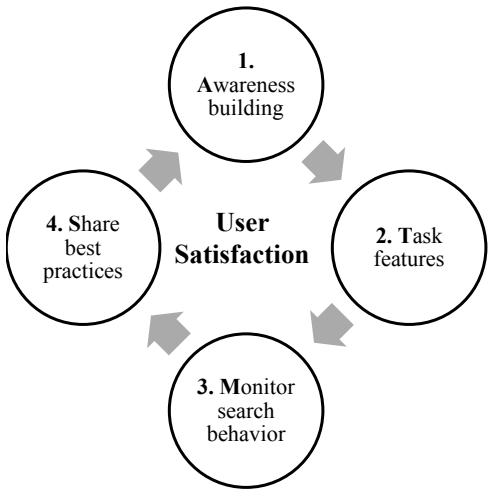

Figure 5.1: The ATMS model to improve the support for complex search in search engines

Users will be supported in two ways. Explicit help will be offered through awareness building and recommendations and implicit support will be given through offering task search features, and sharing of successfully finished search tasks (best practices) as follows:

1. **Awareness:** Our research has shown [99, 100, 101, 103] that many users struggle when they are faced with complex information needs. Especially the comparison between the search behavior and strategies of ordinary Web users [101] with the strategies of search experts [104] shows that experts apply advanced strategies such as tabbed browsing more often than ordinary Web users. In addition users often find it difficult to assess task properties such as difficulty, effort and outcome [100]. I therefore recommend to build awareness among search engine users that not all search needs are the same and that complex search tasks need a different treatment than e.g. look-up tasks and that carrying out complex tasks online might eventually be more effort and take longer than expected. It will also help to teach users already existing techniques (like tabbed browsing) to better support carrying out complex tasks. Of course not all users of Web search engines will equally be able and also willing to adapt their search strategies. My take of the awareness building feature of the ATMS model is outlined in Figure 5.2. Based on the queries that the user has entered over a certain period of time, the system notices that the user has a complex information need. It informs the user that those information needs require different treatment and also offers to turn on the task feature.

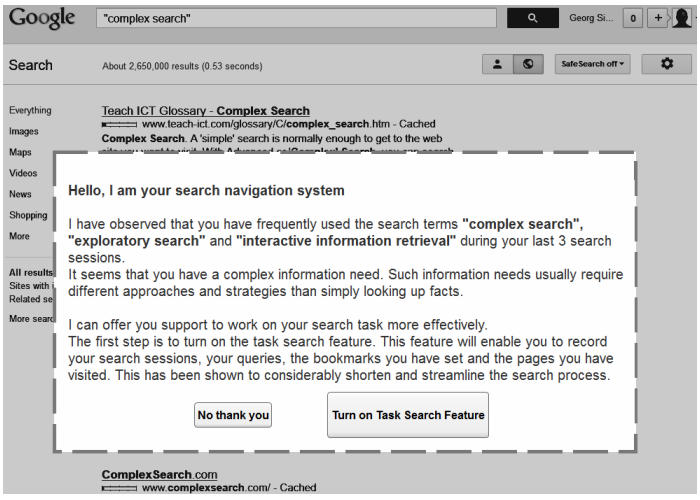

Figure 5.2: Wireframe of the awareness building feature in Google

Researchers in the field of information literacy (compare e.g. [28, 9]) also stress the importance of information skills in the context of information searching. In depth user testing, taking into account information about their Internet user type (a classification of Internet users into information seekers and communication and entertainment seekers derived from people's online activities [54, 103]) will allow segmenting the users according to their willingness to "develop" their search skills.

2. **Task features:** To make the search process more convenient I further recommend to implement a feature in search engines that allows searching in tasks as outlined in Figure 5.3. The results of our user studies confirm that people carry out complex search tasks over longer periods of time [101, 102] - from hours to even weeks. The task option allows users to mark when a search task is being carried out, as opposed to non-task based browsing or querying. At the beginning of such a task the user pushes a "task" button. From now on, all search action is tagged (carrying a tag showing task-specific information). After the task button is pushed, a bar indicating the task option is overlaid on top of the usual search window, and a floating dialog box appears next to the usual search window as illustrated in Figure 5.3 on the next page. This is done in a similar manner as shown by the SearchPad by Bharat [13], but with the slight difference that the SearchPad operated on the query level as opposed to the task level in our case[1]. On the task pad the user can administer task-specific information such as task name, queries used so far, pages bookmarked and content copied. The task can be paused and resumed. Once the task has successfully been carried out it can be stored and retrieved later on.

   Such a task feature fulfills mainly two functions: recording the search process and filtering out certain important parts of the process that the user can reuse later. Those parts are: queries used, Web sites visited, Web sites bookmarked and content copied to the clipboard.

   The task feature has the following advantages: In the case of complex information needs that can go over days or even weeks (such as booking a holiday trip), users can work to a certain point in one session. They can take a break, e.g. talk to family members and discuss options relevant for the search task and can then go on with their search at the point they had left the search session. No opening of browser tabs is needed, no memorizing of the query space is required. All the search details are available at a click.

[1] The SearchPad was nevertheless perceived useful by 150 users in a 4 month user study

It is even thinkable to share the work on the task with other users over social networks, such as family members and friends who will go on the same holiday trip. Apart from this private sharing a public "sharing of best practices" (to be described in Step 4 "Sharing best practices") is thinkable.

Figure 5.3: Wireframe of task search feature in Google

3. Monitor search behavior: Despite building awareness in Step 1, in the case of really complex information needs it can still happen that searchers struggle during their searches. Our research shows that especially older people [99] and people with little Web experience [103] perform significantly worse in the case of complex search tasks. It is therefore advisable to monitor the search behavior of users while carrying out complex search tasks. When erratic or chaotic behavior is identified (like identical queries repeated), searchers get offered, e.g. better suggestions for different queries or related topics as outlined in Figure 5.4. For example if a searcher uses the term "complex search" to find information related to complex search and is struggling with this task, by pushing a button in the side panel the search engine suggests to also use "exploratory search" or "interactive information retrieval" based on the search terms that other

people working on the same task have used. In addition links found and bookmarks set by other people are also offered.

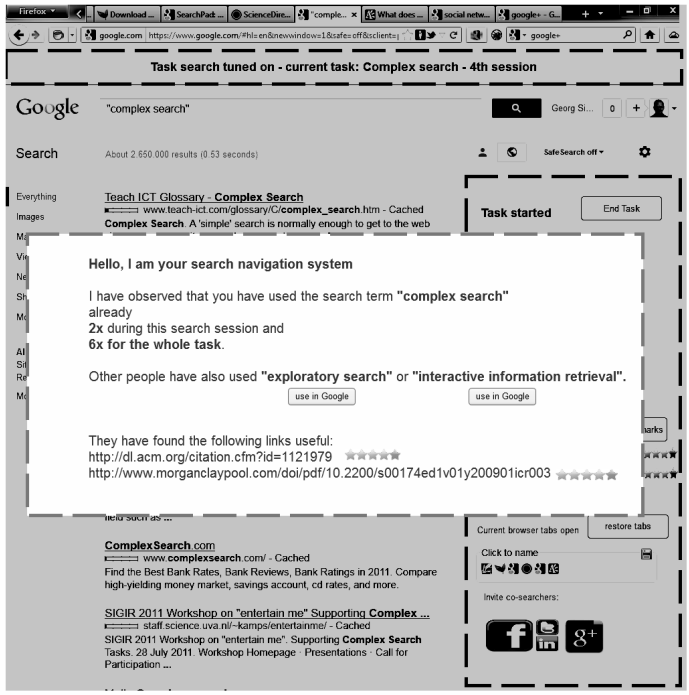

Figure 5.4: Wireframe of search monitor feature in Google

4. **Share best practices:** As the log file that is generated by the SearchLogger [102] contains all relevant information about the search process such as links clicked, Web sites bookmarked and queries issues, this valuable information can easily be re-used. Sharing strategies to tackle those complex search tasks, such as lists of queries that have been useful to other people or sharing the main facets (problem aspects) of the complex search tasks could be used to support other users as illustrated in Figure 5.5. When users push the task button (as explained in Step 2), and specify what task they want to carry out, the search engine automatically notifies the users of existing information relevant for this task. They can check what queries other people have used what sites they have bookmarked and which pages they have visited. This quickly helps them to get an understanding of the main aspects of their complex search need.

The ATMS model as outlined in Figure 5.2 uses the four steps Awareness building, Task based search, Monitoring the search process and Sharing best prac-

tices to improve customer satisfaction. I assume that users will iteratively be going through the ATMS process, each time adding incremental search abilities to their repertoire of search strategies. User satisfaction will automatically grow along with the increased search capabilities.

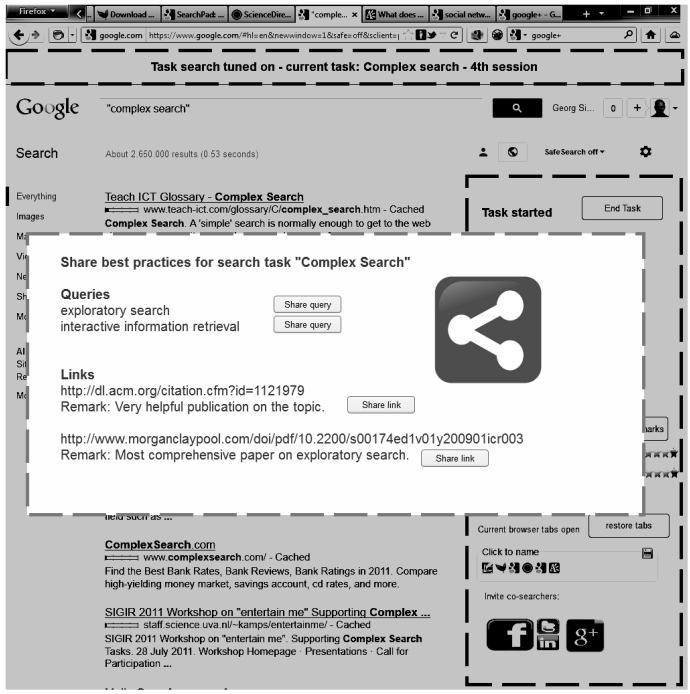

Figure 5.5: Wireframe of task share feature in Google

5.2 Estimation of feature implementation effort

Implementing the ATMS model in a real search engine environment would require different degrees of effort and sophistication for each of the four steps of the ATMS model as outlined in Table 5.1 on the following page.

Educating search engine users that looking for simple facts with a search engine is a different task than e.g. planning a holiday trip would be quite easy to implement by just e.g. monitoring measures like sessions or task time. If users pass certain session or task time thresholds, they could be offered useful information to carry out task searches more efficiently.

Offering the functionality to search in tasks would be slightly more effort. One could just take the code of the current Search-Logger framework [102] and tweak it to let users search in tasks. As the logging functionality is already

there, it would only need some extensions, e.g. mining the log for events such as queries, bookmarks and clicked pages. Those events could then be made accessible to the user as outlined in Figure 5.3. For his master's project at the Institute of Computer Science (University of Tartu) Peeter Jürviste has developed a prototype of an application that lets users record their own search tasks and share those records with other people.

Monitoring the search process and identifying erratic or chaotic behavior is the only functionality that I consider as difficult to implement. It would e.g. require advanced procedures such as machine learning to make the search engine identify erratic search behavior and distinguish it from normal search behavior. In addition the search engine provider would need to have a reasonably big sample of tasks on stock to be able to offer those to other struggling users.

Implementing the functionality to share search results with friends would be straightforward. The important search process information such as queries used or pages bookmarked is extracted and available (done in Step 2). Making this kind of information shareable, e.g. via social networks, is a standard procedure nowadays.

	Effort
Awareness building(Step 1)	easy
Task features (Step 2)	moderate
Monitor search behavior(Step 3)	difficult
Share best practices (Step 4)	easy

Table 5.1: Implementation effort of ATMS in real search engine

In the next section I will discuss the ATMS model in the context of current search challenges.

5.3 Discussion

In this section I will discuss the ATMS model that was defined in the previous section, relate it to present search challenges and initiatives at current Web search engines and also compare it with research on Web information retrieval support systems.

As outlined in Figure 5.6 the challenges in search according to Microsoft lie in mainly three areas:

- increase user satisfaction

- offer context specific help
- implement a way for search engines to be used in sessions not queries
- increase the focus on the task and decision making

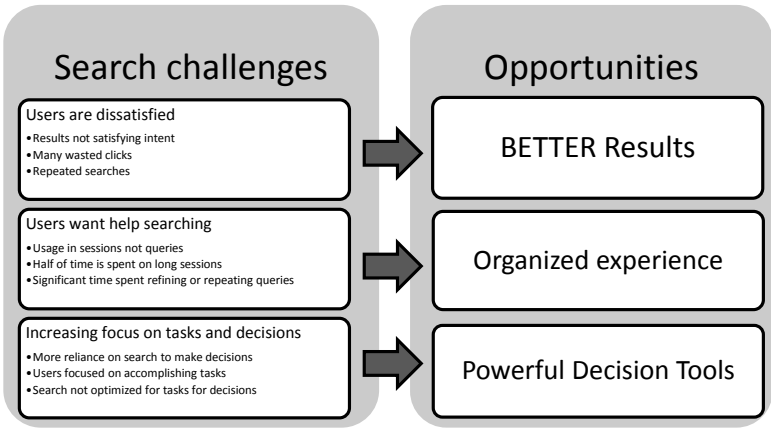

Figure 5.6: Search challenges and opportunities identified by Microsoft [113]

Those challenges and opportunities are interesting, yet Figure 5.6 lacks clear directions about how eventual improvements will be achieved. It is also questionable that "better results" are an opportunity. Mentioning better results as an opportunity again intrinsically assumes that the query-answer method could be improved by increasing the relevance of search results. I assume that the big improvements in complex search are about improving the search process (and not about which search results are ranked highest).

The suggestions that I make in this thesis cover each of the three search challenges (increase user satisfaction, provide help searching, focus on tasks and decisions) that Microsoft has identified. Table 5.2 on page 77 compares Microsoft's search challenges with the four elements of the ATMS model. Not only are all challenges addressed by the ATMS approach, but also additional support is given as follows:

- Search challenge 1 "Users are dissatisfied" identified by Microsoft is supported by Step 1 (Awareness), Step 2 (Tasks) and Step 3 (Monitoring) of the ATMS model:
  - Step 1 of the ATMS model, increasing the awareness among users, will result in fewer repeated searches as users know better what

to expect from a search engine and what not, and automatically decrease users' dissatisfaction. Users will understand that search engines, as they know them, are not the tools to automatically carry out complex search task by pushing a button. Further hints like varying queries more or noting queries to avoid repetitions might further increase user satisfaction by raising the awareness. This will automatically lead to less overall dissatisfaction.

  - Step 2 of the ATMS model, the task search option, will not directly lead to e.g. better search results. Still user satisfaction will rise, as users will be issuing fewer repeated searches.
  - Step 3 of the ATMS model, monitoring the search process and assisting users in case of e.g. erratic behavior will also increase user satisfaction.
  - Step 4 of the ATMS model, share task info, will lead to increased satisfaction, as users will be able to share relevant search results across their search network.

- Search challenge 2 "Provide help searching" identified by Microsoft is supported be Step 2 (Tasks), Step 3 (Monitoring) and Step 4 (Sharing) of the ATMS model:

  - Step 1 of the ATMS model (build awareness) will only indirectly provide help. Users, who are aware of what can be done with current search engines and what not, will less probably fall into certain traps where they need help.
  - Step 2 of the ATMS model, implementing a task search feature (not a session structure as suggested by Microsoft) will automatically also allow the usage of search engines in sessions (and not queries). As I suggest to allow searching in tasks, especially long sessions are automatically supported.
  - Step 3 of the ATMS model, monitoring the search process for indicators of shiftlessness and offering specific help will provide the users with the service level that Microsoft has identified.
  - Step 4 of the ATMS model, sharing task information of already carried out tasks, is a very effective way to help users find exactly what they are looking for. For those tasks users get access to relevant queries, helpful bookmarks and they also are made aware of all the aspects of the search task. This will significantly reduce their search time.

- Search challenge 3 "Better support tasks and decisions" identified by Microsoft is supported be Step 2 (Tasks) and Step 4 (Sharing) of the ATMS model :

  - Step 1 of the ATMS model, build awareness, will not contribute to better supporting search in tasks.
  - Step 2 of the ATMS model, to offer the option to search in tasks will help users who use search engines for making decisions and accomplishing tasks over all.
  - Step 3 of the ATMS model, monitor search process, will not contribute to better supporting search in tasks.
  - Step 4 of the ATMS model, allowing users to share task information will increase the support for decision focused tasks. Especially searches on very complex tasks that require collecting data from various sources, getting to know the problem domain or also understanding inter-dependencies between certain aspect, can be kick-started by learning from how other users have acted in the same situation.

Microsoft search challenges/ATMS model	**Awareness** Step 1	**Tasks** Step 2	**Monitoring** Step 3	**Sharing** Step 4
User dissatisfaction	+	+	+	o
results not satisfying intent	+	o	+	+
wasted clicks	+	+	+	o
repeated searches	+	+	++	o
Better help for search	o	+	+	++
sessions not queries	o	++	+	++
time spent on long sessions	o	++	+	++
time spent refining or repeating queries	o	o	++	++
Tasks and decisions	o	++	o	+
more reliance on search to make decisions	o	+	o	+
users focused on accomplishing tasks	o	++	o	+
search not optimized for tasks and decisions	o	++	o	+

+challenge addressed, o challenge not addressed

Table 5.2: ATMS model versus Microsoft search challenges [113]

Not only Microsoft, but also the other main players in the Web search market are experimenting with advanced search models. According to Gläser et al. [32], Yahoo! has created the "FUSE" model for future search. According to this model, Web search (public information), desktop search (personal information) and search communities (social information) will be more integrated in the future. Yahoo! aims at offering services, which take advantage of the synergies that are being created by this integration. At the beginning of 2012 Google has announced its integration of its core search service with its social network Google+ [109]. This means that Google+ members (but also other people signed in into Google) will be able to choose whether to get search results from the Web only or to also see results from their social network - such as posts on Google+. This way searching across private and public information from one search bar is possible. Google and Yahoo! are pursuing similar approaches here. While they seem to be interesting, they are still focused on look-up needs and do not push any further to better supporting search tasks.

Finally it is interesting to compare the ATMS model with research on Web Information Retrieval Support Systems (WIRSS) [43]. While the WIRSS initiative does not explicitly focus on complex search tasks it tries to improve the support for Web search on a general level. It that context Hoeber [43, p. 3] states activities like "investigating, analyzing, organizing, filtering, understanding, saving, sharing, modifying, manipulating, summarizing" as the main targets for improved support. When comparing those activities with the ATMS model it becomes clear that a considerable overlap can be identified. ATMS supports organizing and filtering, saving and sharing activities. While the ATMS model is clearly designed to support complex search tasks better, the very broad approach of WIRSS seems to be a bit unfocused and not very practical.

Chapter 6

Conclusions and limitations

In this thesis I have examined Web search engines from the angle of improving their support for complex search tasks. Having been an ordinary search engine user myself, I was assuming that "search is search" and Google would find me the answer to all my search needs. The motivation to carry out this research came from my own empirical findings that search engines are incredibly good at something and rather bad at something else. A similar displeasure was expressed by my supervisor about the performance of Google in academia-related searches. Yet it was not clear to me how to judge, what Google was good at and for which activities to lower my expectations.

In this dissertation I have (1) succeeded in analyzing established search disciplines (such as IR, IIR, or QA) and other initiatives (such as exploratory search) and in understanding which are the strengths and limitations of current search tools (RQ1.1). Starting with the exploratory search model, I have developed (2) a generalized model of complex search, comprising the three interactive first level information gathering activities in complex search tasks (aggregation, discovery and synthesis). These three steps are the core activities during complex search tasks and they are relatively well measurable (RQ1.2). This model accounts for fewer aspects than the exploratory search model (the difficult to measure ones such as planning/forecasting or decision making have been omitted).

I have been successful in (3) developing a method to make complex search measurable (RQ2.1). The method is based on a combination of automatically logging the user behavior while searching with a browser add-on and using automatic questionnaires to elicit user specific information such as demographics and satisfaction with task outcome or expected task difficulty. I have presented a number of measures (4), which can be used to characterize complex search behavior - such as the number of sessions it takes to carry out a complex search

task, number of pages viewed, number of queries entered, number of browser tabs opened and query length (RQ2.2 and RQ3.1). We have developed and published (5) a tool called Search-Logger to carry out user studies according to the method (RQ2.3) presented in previous step (3).

In the next step I have presented the results of user studies that we have conducted using the method (3) and the Search-Logger tool (5). The results of a larger user study (6) with ordinary Web users confirm that ordinary people show a search behavior embossed by little strategy and lack of understanding of search engine functionality. They can unexpectedly be struggling with even quite simple tasks (RQ3.1). This is reflected in long search task times and below-expected search success (RQ3.2). I also presented the results of a study with library search experts. As far as recommendations for search engine users carrying out complex search tasks is concerned, this study reconfirms our findings from the study (6) that better performing users show a different search behavior, e.g. they use browser tabs more during their searches - 4.9 tabs in case of library search experts vs. 3.1 tabs in case of worst performing quartile of users in the study (7). In addition the library search experts mainly applied search terms narrowing and search term narrowing and extending strategies in a systematic manner (RQ3.3), while ordinary web users often showed chaotic and erratic search behavior. I showed (8) that being active on the Web in many areas such as communicating online, writing blog articles or commenting on community forums correlate with a person's ability to perform Web search (RQ3.4). The most active Internet user type ranked best during our experiments. The results confirm (9) that people are able to judge difficulty, effort and task outcome for simple tasks. They are significantly less successful when they are asked to do the same for complex search tasks (RQ3.5). We also observed (10) that men and women perform equally well when carrying out complex search tasks (RQ3.6), but (11) age is a significant differentiator regarding search performance, especially for complex tasks (RQ3.7).

Finally I presented the ATMS model (12) to improve the support for complex search tasks in search engines. This model (RQ4.1) comprises the following four elements:

1. Build awareness that complex tasks need different search strategies
2. Offer users the option to carry out their searches in tasks
3. Monitor the search process and offer help when needed
4. Share best practices among users

Of course this model also has its limitations. As ordinary users are mostly unaware of the fact that search engines are tools built for a certain type of

application, it will be a challenge to educate them towards changing their behavior. When limitations regarding e.g. user acceptance are identified, the service level given to users can then be adapted according to the respective user types (RQ4.2). I assume that the ATMS model in its widest sense might not be applicable to all users but rather in varying graduations to different user segments. While advanced users might have been expecting this task feature for a long time, less search-savvy searchers would just need basic support at the beginning. A feasible way to segment users would be eliciting their Internet user types [103], deriving their search abilities and offer Internet user type specific assistance.

Closing words

In this PhD project I have produced a number of significant results to better support complex search with search engines. Based on these results a number of follow-up projects were initiated.

At the University of Tartu (Institute of Computer Science) the PhD student Dmitri Danilov bases his research on the results achieved in this dissertation by trying to improve the support for users struggling with complex search tasks. He builds a technology to help users explore the search result space by automatically offering them additional queries based on the search results they have identified as relevant already. This feature fits well into Step 3 (Monitor search behavior) of the ATMS model as outlined in Figure 5.1 on page 68. Once a search engine identifies erratic search behavior (such as a user entering the same query multiple times), this technology would immediately offer valuable help.

The master student Peeter Jürviste has developed a proxy-based solution for the Search-Logger. This will make the installation of the Search-Logger being easier as the proxy-based version will not depend on a certain version of the Firefox browser installed at the computers that the study participants work with. In addition, the proxy-based version will get rid of the following problem that we discovered just before the user study in Hamburg: During the roll-out of the instant predictions feature (search results are changed instantaneously according to the queries typed into the search bar) Google had switched from loading a new web page after a user had entered a query to building the SERPs with Java script. We had to turn off this functionality at all computers during the Hamburg user study and later turn it on again.

At the University of Applied Sciences in Hamburg Prof. Dirk Lewandowski plans to merge the Search-Logger [102] with the Relevance Assessment Tool by Lewandowski and Sünkler [68]. According to Prof. Lewandowski, adding the

relevance dimension (on the search result level) to the Search-Logger will significantly enhance the technological support to carry out search engine related user studies.

In future work it would be interesting to investigate the ability to measure core exploratory search concepts like learning, planning, and decision making. We are also playing with the idea to run experiments with the mobile version of the Search-Logger (developed at the University of Tartu by the bachelor student Gleb Štšenov). This mobile version extends a real browser with logging features on a mobile phone.

This dissertation has greatly succeeded in making complex search measurable. Through the creation of the Search-Logger experimentation framework and the user studies, complex search user behavior has been analyzed in detail and potential directions for improved search engine support have been given in the form of the ATMS model.

Bibliography

[1] Online extra: Google's goal: "Understand everything", May 2004. Available from: `http://www.businessweek.com/magazine/content/04_18/b3881010_mz001.htm` [cited 2012-01-26].

[2] Web search query, July 2012. Available from: `http://itlaw.wikia.com/wiki/Web_search_query` [cited 2012-07-03].

[3] J. Allan. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proceedings of the 14th Text Retrieval Conference (TREC 2005)*, Gaithersburg, MD, USA, 2005. NIST.

[4] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 433–440, Salvador, Brazil, 2005. ACM New York, NY, USA.

[5] A. Aula. User study on older adults' use of the web and search engines. *Universal Access in the Information Society*, 4(1):67–81, 2005.

[6] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern Information Retrieval*. Addison-Wesley New York, 1999.

[7] H. Balinsky, A. Balinsky, and S. J. Simske. Automatic text summarization and small-world networks. In *Proceedings of the 11th ACM symposium on Document engineering*, pages 175–184, Mountain View, CA, USA, 2011. ACM New York, NY, USA.

[8] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, volume 17, pages 10–17, Madrid, Spain, 1997.

[9] D. Bawden and L. Robinson. The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2):180–191, Apr. 2009.

[10] N. J. Belkin. A methodology for taking account of user tasks, goals and behavior for design of computerized library catalogs. *ACM SIGCHI Bulletin*, 23(1):61–65, Jan. 1991.

[11] N. J. Belkin, R. Oddy, and H. Brooks. Ask for Information Retrieval: Part I. Background and Theory. *Journal of Documentation*, 38(2):61–71, Dec. 1982.

[12] D. Bell and I. Ruthven. Searcher's assessments of task complexity for web searching. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR)*, pages 57–71, Berlin/Heidelberg, Germany, 2004. Springer.

[13] K. Bharat. SearchPad: Explicit capture of search context to support web search. *Computer Networks*, 33(1-6):493–501, 2000.

[14] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), Apr. 2003. Available from: `http://informationr.net/ir/8-3/paper152.html` [cited 2012-03-04].

[15] B. Boyce, C. Meadow, and D. Kraft. *Measurement in information science.* Academic Press, San Diego, CA, USA, 1994.

[16] R. Capra. HCI browser: A tool for studying web search behavior. In *Proceedings of the American Society for Information Science and Technology*, volume 47, pages 1–2. Wiley Online Library, 2010.

[17] C. N. Carlson. Information overload, retrieval strategies and Internet user empowerment, 2003. Available from: `http://eprints.rclis.org/2248/1/Information_Overload.pdf` [cited 2012-01-26].

[18] I. Ceaparu, J. Lazar, K. Bessiere, J. Robinson, and B. Shneiderman. Determining causes and severity of End-User frustration. *International Journal of Human-Computer Interaction*, 17(3):333–356, 2004.

[19] J. Cheng, X. Hu, and P. B. Heidorn. New measures for the evaluation of interactive information retrieval systems: Normalized task completion time and normalized user effectiveness. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–9, Nov. 2010.

[20] J. Chin and W. Fu. Interactive effects of age and interface differences on search strategies and performance. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 403–412, Atlanta, GA, USA, 2010. ACM New York, NY, USA.

[21] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pages 33–40, Santa Fe, NM, USA, 2001. ACM New York, NY, USA.

[22] C. W. Cleverdon and E. M. Keen. Factors determining the performance of indexing systems (2 volumes). Technical report, Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK, 1966. Available from: `http://www.sigir.org/museum/pdfs/ASLIB%20CRANFIELD%20RESEARCH%20PROJECT-1960/pdfs/frontmatter.pdf` [cited 2012-07-24].

[23] P. Clough and B. Berendt. Report on the TrebleCLEF query log analysis workshop 2009. *ACM SIGIR Forum*, 43(2):71–77, 2009.

[24] M. Czerwinski, E. Horvitz, and S. Wilhite. A diary study of task switching and interruptions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 175–182, Vienna, Austria, 2004. ACM New York, NY, USA.

[25] H. T. Dang, J. Lin, and D. Kelly. Overview of the TREC 2006 question answering track. In *15th Text Retrieval Conference*, Gaithersburg, MD, USA, 2006.

[26] A. Dickinson, M. J. Smith, J. L. Arnott, A. F. Newell, and R. L. Hill. Approaches to web search and navigation for older computer novices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, pages 281–290, San Jose, CA, USA, 2007. ACM New York, NY, USA.

[27] S. T. Dumais and N. J. Belkin. The TREC interactive track: Putting the user into search. In *TREC Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, USA, 2005.

[28] M. B. Eisenberg. Information literacy: Essential skills for the information age. *DESIDOC Journal of Library & Information Technology*, 28(2):39–47, Mar. 2010.

[29] K. El-Arini and C. Guestrin. Beyond keyword search: discovering relevant scientific literature. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, pages 439–474, San Diego, California, USA, 2011. ACM New York, NY, USA.

[30] S. El Ayari and B. Grau. A framework of evaluation for Question-Answering systems. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 744–748. Springer, Berlin/Heidelberg, Germany, 2009.

[31] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2):147–168, 2005.

[32] V. Gläser, J. Eberspächer, and S. Holtel. Suchen und Finden als Bindeglied zum Produktportfolio. In *Suchen und Finden im Internet*, pages 102–107. Springer, Berlin/Heidelberg, Germany, 2007.

[33] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, New Orleans, LA, USA, 2001. ACM New York, NY, USA.

[34] M. A. Greenwood. *Open-Domain Question Answering*. PhD thesis, University of Sheffield, Sheffield, UK, 2005. Available from: `http://nlp.shef.ac.uk/Completed_PhD_Projects/Greenwood2006.pdf` [cited 2012-07-20].

[35] J. Gwizdka. Assessing cognitive load on web search tasks. *The Ergonomics Open Journal*, 2:114–123, 2009.

[36] J. Gwizdka and I. Spence. What can searching behavior tell us about the difficulty of information tasks? A study of web navigation. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–22, Jan. 2006.

[37] B. Hachey, G. Murray, and D. Reitter. The embra system at duc 2005: Query-oriented multi-document summarization with a very large latent semantic space. In *Proceedings of the Document Understanding Conference (DUC) 2005, Vancouver, BC, Canada*, 2005.

[38] A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 362–370, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[39] N. Höchstötter and M. Koch. Standard parameters for searching behaviour in search engines and their empirical evaluation. *Journal of Information Science*, 35(1):45–65, Feb. 2009.

[40] A. J. Head and M. B. Eisenberg. How college students use the web to conduct everyday life research. *First Monday; Volume 16, Number 4 - 4 April 2011*, 2011.

[41] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 195–206, Stanford, CA, USA, 2008. ACM New York, NY, USA.

[42] C. Hölscher and G. Strube. Web search behavior of internet experts and newbies. *Computer Networks*, 33(1-6):337–346, June 2000.

[43] O. Hoeber. Web information retrieval support systems: The future of web search. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, volume 3, pages 29–32, Los Alamitos, CA, USA, 2008. IEEE Computer Society.

[44] M. E. Hupfer and B. Detlor. Gender and web information seeking: A self-concept orientation model. *Journal of the American Society for Information Science and Technology*, 57(8):1105–1115, 2006.

[45] P. Ingwersen. *Information retrieval interaction*. Taylor Graham Publishing, London, UK, 1992.

[46] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, Berlin/Heidelberg, Germany, 1 edition, Aug. 2005.

[47] L. A. Jackson, K. S. Ervin, P. D. Gardner, and N. Schmitt. Gender and the internet: Women communicating and men searching. *Sex Roles*, 44(5):363–379, 2001.

[48] B. J. Jansen, R. Ramadoss, M. Zhang, and N. Zang. Wrapper: An application for evaluating exploratory searching outside of the lab. In

Proc. SIGIR 2006: Workshop on Evaluating Exploratory Search Systems, pages 14–19, Seattle, WA, 2006. ACM New York, NY, USA.

[49] B. J. Jansen and A. Spink. How are we searching the world wide web? a comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263, 2006.

[50] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: a study of user queries on the web. *SIGIR Forum*, 32(1):5–17, Apr. 1998.

[51] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on web search engines. *Journal of the American Society for Information Science and Technology*, 58(6):862–871, Apr. 2007.

[52] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2):207–227, Mar. 2000.

[53] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 154–161, Salvador, Brazil, 2005. ACM New York, NY, USA.

[54] V. Kalmus, P. Pruulmann-Vengerfeldt, and M. Keller. Quality of life in a consumer and information society: defining consumer and information society. *2008 Estonian Human Development Report*, pages 102–103, 2009.

[55] M. Kellar, C. Watters, and M. Shepherd. A field study characterizing web based information seeking tasks. *Journal of the American Society for Information Science and Technology*, 58(7):999–1018, May 2007.

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

[57] D. Kelly, S. Dumais, and J. Pedersen. Evaluation challenges and directions for information seeking support systems. *IEEE Computer*, 42(3):60–66, 2009.

[58] A. Kerne and S. M. Smith. The information discovery framework. In *Proceedings of the 5th conference on Designing interactive systems: Processes, practices, methods, and techniques*, pages 357–360, Cambridge, MA, USA, 2004. ACM New York, NY, USA.

[59] E. Koh, A. Kerne, and R. Hill. Creativity support: information discovery and exploratory search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 895–896, Amsterdam, The Netherlands, 2007. ACM New York, NY, USA.

[60] W. Kraaij and W. Post. Task based evaluation of exploratory search systems. In *Workshop on Evaluating Exploratory Search Systems at the 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR '06)*, pages 24–27, Seattle, WA, 2006. ACM New York, NY, USA.

[61] B. Krause, A. Hotho, and G. Stumme. A comparison of social bookmarking with traditional search. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 101–113, Berlin/Heidelberg, Germany, 2008. Springer.

[62] J. E. Kubeck. Finding information on the world wide web: Exploring older adults' exploration. *Educational Gerontology*, 25(2):167–183, 1999.

[63] B. Kules and R. Capra. Designing exploratory search tasks for user studies of information seeking support systems. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 419–420, Austin, TX, USA, 2009. ACM.

[64] L. Langford. Surf's up: harnessing information overload. *Engineering Management Review, IEEE*, 38(1):164–165, 2010.

[65] D. Lewandowski. The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, 64(6):915–937, 2008.

[66] D. Lewandowski. *Web Search Engine Research*. Emerald Group Publishing, Bingley, UK, 2012.

[67] D. Lewandowski and N. Höchstötter. Web searching: A quality measurement perspective. *Web Search, Information Science and Knowledge Management*, 14:309–340, 2008.

[68] D. Lewandowski and S. Sünkler. Relevance assessment tool: Ein werkzeug zum design von retrievaltests sowie zur weitgehend automatisierten erfassung, aufbereitung und auswertung der daten. In *Proceedings der 2. DGI-Konferenz: Social Media und Web Science - Das Web als Lebensraum.*, pages 237–249, Frankfurt am Main, 2012. DGI.

[69] Y. Li. *Relationships among work tasks, search tasks, and interactive information searching behavior*. PhD thesis, Rutgers - The State University of New Jersey, New Brunswick, NJ, USA, 2008. Available from: `http://proquest.umi.com/pqdweb?did=1683299241&sid=1&Fmt=2&clientId=79356&RQT=309&VName=PQD` [cited 2012-03-05].

[70] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, Nov. 2008.

[71] Y. Li, Y. Chen, J. Liu, Y. Cheng, X. Wang, P. Chen, and Q. Wang. Measuring task complexity in information search from user's perspective. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–8, 2011.

[72] J. Liu. Personalizing information retrieval using task features, topic knowledge, and task products. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 855–855, Boston, MA, USA, 2009. ACM New York, NY, USA.

[73] Z. Liu and X. Huang. Gender differences in the online reading environment. *Journal of Documentation*, 64(4):616–626, July 2008.

[74] S. Livingstone. Media literacy and the challenge of new information and communication technologies. *The Communication Review*, 7(1):3–14, 2004.

[75] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay. The influence of task and gender on search and evaluation behavior using google. *Information Processing and Management*, 42(4):1123–1131, 2006.

[76] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, Apr. 1958.

[77] M. Machill, C. Neuberger, W. Schweiger, and W. Wirth. Navigating the internet: A study of german-language search engines. *European Journal of Communication*, 19(3):321–347, Aug. 2004.

[78] G. Marchionini. *Information seeking in electronic environments*. Cambridge University Press, Cambridge, MA, USA, Mar. 1997.

[79] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, Volume 49(4):41–46, 2006.

[80] K. Markey. Twenty five years of end user searching, part 1: Research findings. *Journal of the American Society for Information Science and Technology*, 58(8):1071–1081, June 2007.

[81] A. H. Maslow, R. Frager, and J. Fadiman. *Motivation and personality.* HarperCollins Publishers, New York, NY, USA, 3 edition, 1987.

[82] B. Meyer, R. A. Sit, V. A. Spaulding, S. E. Mead, and N. Walker. Age group differences in world wide web navigation. In *CHI '97 extended abstracts on Human factors in computing systems: looking to the future*, CHI EA '97, pages 295–296, Atlanta, GA, USA, 1997. ACM New York, NY, USA.

[83] D. Moldovan, M. Paşca, S. Harabagiu, and M. Surdeanu. Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS)*, 21(2):133–154, Apr. 2003.

[84] R. W. Morrell, C. B. Mayhorn, and J. Bennett. A survey of world wide web use in Middle-Aged and older adults. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 42(2):175–182, June 2000.

[85] R. Navarro-Prieto, M. Scaife, and Y. Rogers. Cognitive strategies in web searching. In *Proceedings of the 5th Conference on Human Factors & the Web*, pages 43–56, Austin, TX, USA, 1999. NIST, USA.

[86] A. Nenkova, S. Maskey, and Y. Liu. Automatic summarization, 2011. Available from: `http://aclweb.org/anthology/P/P11/P11-5003.pdf` [cited 2011-11-07].

[87] A. Nenkova and R. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL*, pages 145–152, Boston, MA, USA, 2004. Association for Computational Linguistics.

[88] P. Pruulmann-Vengerfeldt. *Information technology users and uses within the different layers of the information environment in Estonia.* PhD thesis, University of Tartu, Tartu, Estonia, 2006. Available from: `http://dspace.utlib.ee/dspace/handle/10062/173` [cited 2011-12-08].

[89] R. W. Reeder, P. Pirolli, and S. K. Card. Webeyemapper and weblogger: Tools for analyzing eye tracking data collected in web-use studies. In *CHI'01 extended abstracts on Human factors in computing systems*, pages 19–20, Seattle, WA, USA, 2001. ACM New York, NY, USA.

[90] J. C. Reinard. *Communication Research Statistics.* Sage Publications, Inc., Thousand Oaks, CA, USA, Apr. 2006.

[91] D. Robins. Interactive information retrieval: Context and basic notions. *Informing Science Journal*, 3(2):57–62, 2000.

[92] M. Roy and M. T. Chi. Gender differences in patterns of searching the web. *Journal of Educational Computing Research*, 29(3):335–348, 2003.

[93] P. Runnel. *The transformation of the Internet usage practices in Estonia.* Tartu University Press, Tartu, Estonia, 2009.

[94] I. Ruthven. Interactive information retrieval. *Annual Review of Information Science and Technology*, 42(1):43–91, Jan. 2008.

[95] J. Sauro and J. R. Lewis. *Quantifying the User Experience: Practical Statistics for User Research.* Morgan Kaufmann Publishers Inc., Waltham, MA, USA, May 2012.

[96] B. Schiffman, A. Nenkova, and K. McKeown. Experiments in multidocument summarization. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 52–58, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc., USA.

[97] B. Shneiderman. Designing information-abundant web sites: issues and recommendations. *International Journal of Human Computer Studies*, 47:5–30, 1997.

[98] G. Singer, U. Norbisrath, and D. Danilov. Complex search: Aggregation, discovery, and synthesis. *Proceedings of the Estonian Academy of Sciences*, 61(2):89–106, 2012.

[99] G. Singer, U. Norbisrath, and D. Lewandowski. Impact of gender and age on performing complex search tasks online. In *Mensch & Computer 2012*, Konstanz, Germany, Sept. 2012. Oldenbourg Wissenschaftsverlag Munich - in press. Available from: `http://arxiv.org/abs/1206.1494` [cited 2012-06-21].

[100] G. Singer, U. Norbisrath, and D. Lewandowski. Ordinary search engine users assessing difficulty, effort, and outcome for simple and complex search tasks. In *4th Information Interaction in Context Symposium iiiX 2012*, Nijmegen, Netherlands, Aug. 2012. ACM New York - in press. Available from: `http://arxiv.org/abs/1206.2528` [cited 2012-06-15].

[101] G. Singer, U. Norbisrath, and D. Lewandowski. Ordinary search engine users carrying out complex search tasks, 2012. Manuscript submitted for publication. Available from: `http://arxiv.org/abs/1206.1492` [cited 2012-03-15].

[102] G. Singer, U. Norbisrath, E. Vainikko, H. Kikkas, and D. Lewandowski. Search-Logger - Analyzing exploratory search tasks. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 751–756, Taichung, Taiwan, 2011. ACM New York, NY, USA.

[103] G. Singer, P. Pruulmann-Vengerfeldt, U. Norbisrath, and D. Lewandowski. The relationship between internet user type and user performance when carrying out simple vs. complex search tasks. *First Monday*, 17(6), June 2012. Available from: `http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3960/3245` [cited 2012-06-11].

[104] K. Singer, G. Singer, K. Lepik, U. Norbisrath, and P. Pruulmann-Vengerfeldt. Search strategies of library search experts. In *Conference on Qualitative and Quantitative Methods in Libraries QQML 2011*, Athens, Greece. in press. Available from: `http://arxiv.org/abs/1206.2465` [cited 2012-06-18].

[105] J. Steinberger and K. Jezek. Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of ISIM 2004*, pages 93–100, Roznos pod Radhostem, Czech Republic, Apr. 2004.

[106] B. C. Taylor, M. Mayer, and O. Buyukkokten. Interface for a universal search engine, July 2008. Available from: `http://www.freepatentsonline.com/EP1700239.html` [cited 2011-05-20].

[107] A. Thatcher. Web search strategies: The influence of web experience and task type. *Information Processing & Management*, 44(3):1308–1329, May 2008.

[108] D. Turnbull. Webtracker: A tool for understanding web use, 1998. Available from: `http://www.ischool.utexas.edu/~donturn/research/webtracker/` [cited 2012-03-07].

[109] L. Ulanov. Google merges search and Google+ into social media juggernaut. Available from: `http://mashable.com/2012/01/10/google-launches-social-search/` [cited 2012-04-03].

[110] P. Vakkari and S. Huuskonen. Search effort degrades search output but improves task outcome. *Journal of the American Society for Information Science and Technology*, 63(4):657–670, Apr. 2012.

[111] E. Voorhees and D. M. Tice. The TREC-8 question answering track report. In *Text Retrieval Conference TREC-8*, volume 8, pages 77–82. NIST, USA, 1999.

[112] E. M. Voorhees and D. K. Harman. *TREC: Experiment and evaluation in information retrieval*. MIT Press, Boston, MA, USA, 2005.

[113] S. Weitz. Search isn't search. Microsoft company report, Search Marketing Expo SMX, Munich, Germany, 2009.

[114] M. D. White and M. Iivonen. Assessing level of difficulty in web search questions. *The Library Quarterly*, 72(2):205–233, Apr. 2002.

[115] R. White and R. Roth. *Exploratory Search*. Morgan & Claypool Publishers, San Rafael, CA, USA, 2009.

[116] R. W. White, G. Muresan, and G. Marchionini. Report on ACM SIGIR 2006 workshop on evaluating exploratory search systems. *ACM SIGIR Forum*, 40(2):52–60, Dec. 2006.

[117] R. W. White and R. A. Roth. Exploratory search: Beyond the Query-Response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[118] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 155–162, Singapore, 2008. ACM New York, NY, USA.

Acknowledgments

A big thank you to everybody who supported me during my PhD studies. I would especially like to thank my great supervisory team Dr. Ulrich Norbisrath, for his great and inspiring vision to tackle a real Web search problem, Prof. Dirk Lewandowski for his guidance and mentoring in all user study and search engine evaluation related aspects and Prof. Eero Vainikko for his general support, help and advice. In addition I would like to thank Hannu Kikkas for his help on the implementation of the Search-Logger tool, and Natalie Ivanova for her help with the data analysis of the Hamburg study.

A big thank you also goes to my wife Kristiina for her support all along the way, during long work days and weekends. Kristiina has always managed listening to me and her advice has helped me a great deal, more than once during the last three years.

Finally I would like to acknowledge that the research presented in this dissertation was partly supported by the European Union Social Fund and by the Estonian Information Technology Foundation (EITSA) and the Tiger University program.

Appendix

Appendix A

Technical implementation details

In this chapter I describe the details about how the technical implementation of the Search-Logger study framework to carry out user studies with complex search tasks was realized (for details about the method in general please refer to Section 3.2 on page 29). The framework comprises two parts: One part is the Search-Logger to automatically record the user events during usability studies and administering automatic questionnaires and the second part is the statistical analyzer of the log file. First, I describe the sources of data and the data collection process. Then I outline the development work related to the logging part of the Search-Logger followed by the summary of the Search-Logger Analyzer implementation.

A.1 Sources of data and data collection process

The Search-Logger needs to be installed at computers that study participants use during the experiment. The Search-Logger works with two kinds of data.

1. standard user events such as links clicked, browser tabs opened or queries entered are automatically recorded
2. user feedback is gathered through automatic questionnaires at the beginning of the experiment and before and after each search task

The process how users are guided through the experiment and what data is collected is illustrated in Figure A.1 on the following page. When the users begin the experiment, they also manually start the Search-Logger. They are

then asked to fill in a demographic form that collects data such as gender, age, and information related to the users' Internet habits.

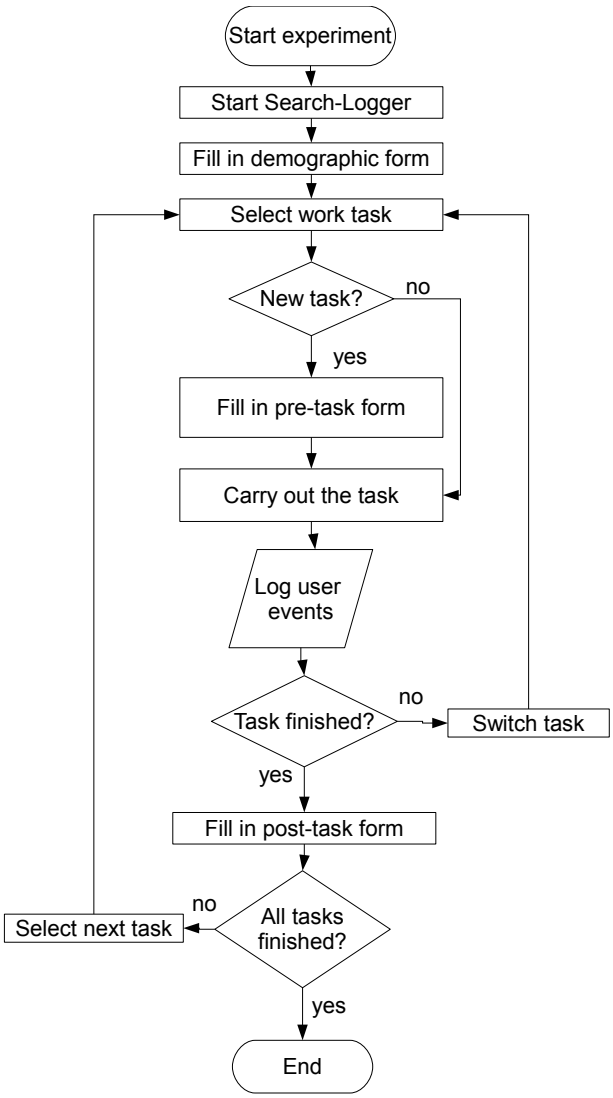

Figure A.1: Flow chart of user process

Once this form is submitted, the users select the first work task followed by a couple of task specific questions to gather information such as the users' assessment of the difficulty of the task. The pre-task form is only presented once depending on whether this task was previously started or not. Then

the users carry out the task or continue working on it if it has previously been started. When they are done, they push the "finish the task" button of the Search-Logger and are then asked to fill in a post-task form. This form gathers information such as how difficult this task was perceived. The study participants then choose the next task and proceed in the same manner. They can always pause a task and switch to another one if they want. When they have carried out all tasks, they can finish the experiment by pushing the finish button of the Search-Logger.

A.2 Search-Logger

The Search-Logger [102] is a tool to evaluate the user behavior when carrying out complex search tasks. I will first name the requirements that we identified in advance of the development work, then I summarize the implementation process itself.

Requirements

As already stated in Section 3.2, where I described the method that we decided to use to carry out our experiments, we found the following requirements to be important for the development of the Search-Logger recording tool:

- We should have the possibility to administer tasks to users taking part in the experiment
- We needed to be able to automatically record specific user action such as links clicked, browser tabs opened, queries entered
- Each user event should be tagged with task number, time, date and user number
- The implementation should be done in a way that the users' normal search behavior is not disturbed
- The record of the user action should be gathered in a single central database
- The implementation of the recording tool should be deploy-able across many operating systems such as Windows, Linux and Mac
- The approach should be low cost, i.e. using proprietary software should be avoided

- The implementation should be easy to install so that study participants could get it running remotely by reading a simple instruction
- Users should be able to start, pause and stop search tasks whenever they wanted
- Users should be able to switch between tasks whenever they wanted
- We needed to be able to gather demographic information at the beginning of the experiment and task specific information before and after each search tasks

Implementation

Considering above requirements, the Search-Logger recording tool is realized as a browser plug-in (developed in Java-Script) for Firefox, combined with a remote log storage database and an analysis environment as outlined in Figure A.2. It fulfills the following three main tasks: (i) administers pre-compiled search tasks to users, (ii) gathers implicit information about the search process by automatically logging various browser events as outlined in the next paragraph, (iii) gathers explicit user feedback via standardized questionnaires supplied before and after each search task. Each logged event is tagged with task specific information such as task name, task number, user number, and a time stamp. Based on this information the task performance can be analyzed and evaluated.

All data is centrally collected at a dedicated server. We log the search process by gathering data on all measurable standard user events such as total search time, number of Web pages visited, number of browser tabs opened, search queries entered and number of search sessions started and ended.

Developing the Search-Logger took about 1 year (net development time was approximately 900 - 1000 man hours).

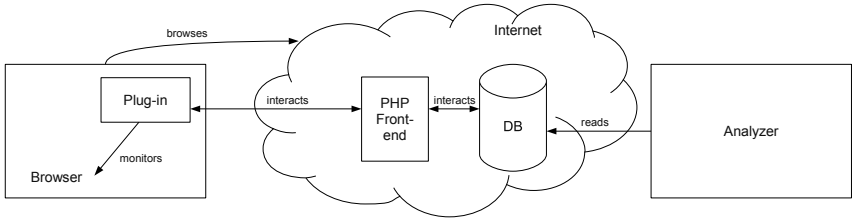

Figure A.2: Search-Logger architecture

A.3 Log Analyzer

I will describe the technical implementation of the Log Analyzer by means of the larger user study conducted in Hamburg in July 2011 (as introduced in Section 4.3 on page 45). As already stated in the data collection section above, we collected two types of data in the course of this study: On the one hand the user behavior was automatically logged (and the data stored into a log file), on the other hand we asked the study participants to fill in automatic questionnaires at the beginning of the experiment and before and after each search task.

Requirements

The requirements for the Log Analyzer were less stringent than for the recording part of the Search-Logger.

- The raw log file (comma separated values) consisted of about 30 000 lines. Hence an appropriate method to process it efficiently was needed
- The analyzer needed to be able to combine both the log data as well as the information gathered through the automatic questionnaires
- The final output of the analyzer should be the measures that we needed for our publications
- As in the case of the recording part, non standard software (such as SPSS) should be avoided to keep costs of the PhD project down

Implementation

The realization of the statistical analyzer was insofar more difficult than expected, as especially the log-file created at our Hamburg study consisted of close to 30 000 log entries (as outlined in Figure A.3). After starting to work on the data analysis part, I quickly realized that I would run into performance issues quite soon.

78	2	http://googleads.g.doubleclick.net/pagead/drt/s;;	141.22.170.152	1.313E+09	40765.4742	2_Aachen_Temp_(KERNAUFGABE)_2	
78	2	http://www.kinkaa.de/wetter/Aachen_Nordrhein-Westfalen_Deutschland.4179.htn	141.22.170.152	1.313E+09	40765.4742	2_Aachen_Temp_(KERNAUFGABE)_2	
78	2	https://googleads.g.doubleclick.net/pagead/drt/si?p=CAA&ut=AFAKxlQAAAAATkJD\	141.22.170.152	1.313E+09	40765.4742	2_Aachen_Temp_(KERNAUFGABE)_2	
79	2	about:blank;;	141.22.170.151	1.313E+09	40765.4742	1_Komponist_(KERNAUFGABE)_1	
79	2	about:blank;;	141.22.170.151	1.313E+09	40765.4742	1_Komponist_(KERNAUFGABE)_1	
79	1	http://www.google.de/search?q=+zauberfl%C3%B6te+&ie=utf-8&oe=utf-8&aq=t&rl	141.22.170.151	1.313E+09	40765.4742	1_Komponist_(KERNAUFGABE)_1	
72	2	about:blank;;	141.22.170.158	1.313E+09	40765.4742	1_Komponist_(KERNAUFGABE)_1	
72	1	http://www.google.de/search?hl=de&client=firefox-a&hs=ngP&rls=org.mozilla%3Ac	141.22.170.158	1.313E+09	40765.4742	1_Komponist_(KERNAUFGABE)_1	
72	1	http://www.google.de/advanced_search?q=zauberfl%C3%B6te&hl=de&client=firefc	141.22.170.158	1.313E+09	40765.4743	1_Komponist_(KERNAUFGABE)_1	
84	14	IF(a=(35))RB(b=(2))RB(d=(3))RB(e=(1))RB(k=(4))RB(l=(5))RB(m=(4))RB(n=(4))RB(o=(4)	141.22.170.192	1.313E+09	40765.4743		
84	2	file:///C:/Users/ntadmin/AppData/Roaming/Mozilla/Firefox/Profiles/s32fa9zp.defau	141.22.170.192	1.313E+09	40765.4743	1_Komponist_(KERNAUFGABE)_1	
72	2	about:blank;;	141.22.170.158	1.313E+09	40765.4743	1_Komponist_(KERNAUFGABE)_1	
72	1	http://www.google.de/search?q=zauberfl%C3%B6te+%22komponist%22&hl=de&cli	141.22.170.158	1.313E+09	40765.4743	1_Komponist_(KERNAUFGABE)_1	
82	2	http://googleads.g.doubleclick.net/pagead/drt/s;;	141.22.170.191	1.313E+09	40765.4744	1_Komponist_(KERNAUFGABE)_1	
82	1	http://googleads.g.doubleclick.net/pagead/ads?client=ca-pub-6506748959449825&	141.22.170.191	1.313E+09	40765.4744	1_Komponist_(KERNAUFGABE)_1	
82	2	http://www.salzburg-rundgang.at/geboren_gelebt/wolfgang_amadeus_mozart/;;	141.22.170.191	1.313E+09	40765.4744	1_Komponist_(KERNAUFGABE)_1	
82	2	http://www.salzburg-rundgang.at/geboren_gelebt/wolfgang_amadeus_mozart/;;W	141.22.170.191	1.313E+09	40765.4744	1_Komponist_(KERNAUFGABE)_1	
82	2	https://googleads.g.doubleclick.net/pagead/drt/si?p=CAA&ut=AFAKxlQAAAAATkJD\	141.22.170.191	1.313E+09	40765.4744	1_Komponist_(KERNAUFGABE)_1	
82	3	Clipboard%20change%20detected%20...;;	141.22.170.191	1.313E+09	40765.4745	1_Komponist_(KERNAUFGABE)_1	
82	4	Clipboard%20contents;;27.01.1756	141.22.170.191	1.313E+09	40765.4745	1_Komponist_(KERNAUFGABE)_1	
76	11	Displaying%20post-SC%20form%20for%20SC%20index%200;;User%20has%20presur	141.22.170.149	1.313E+09	40765.4745	1_Komponist_(KERNAUFGABE)_1	
76	25	User%20opened%20a%20tab;;	141.22.170.149	1.313E+09	40765.4745	1_Komponist_(KERNAUFGABE)_1	
76	2	file:///C:/Users/ntadmin/AppData/Roaming/Mozilla/Firefox/Profiles/5vsnz9wc.defa	141.22.170.149	1.313E+09	40765.4745	1_Komponist_(KERNAUFGABE)_1	

Figure A.3: Raw log of Hamburg Study

Generic solution available? At the beginning of the data analysis efforts I invested a lot of time into finding a generic solution and checking the packages regarding their suitability for our needs. I experimented with a couple of open source packages like "GanttProject" (www.ganttproject.biz), packages for Gnuplot or SIMILE Widgets (www.simile-widgets.org). Unfortunately none of the packages that I tried really offered the specific functionality that I needed. For example SIMILE Widgets offered a very nice way to display timed events by using a time line representation as outlined in Figure A.4. An approach like this was interesting for illustration purposes, but I needed to analyze search specific measures like the time users spent on SERPs, number of browser tabs opened and all kinds of query reformulations. To analyze those very search specific measures, none of the out of the box solutions was really helpful. I therefore opted for implementing the analyzer myself.

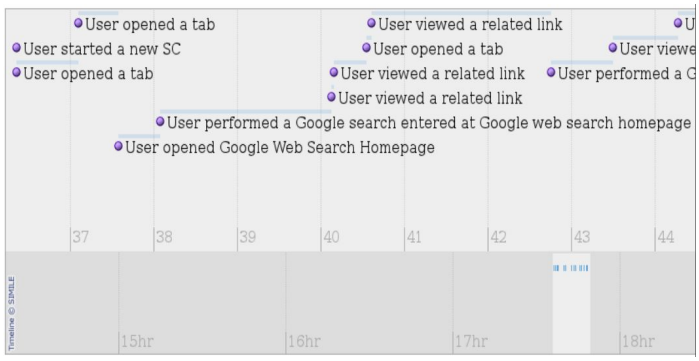

Figure A.4: Time line representation of user action with SIMILE Data Visualization Web Widget

My implementation I decided not to go for an out of the box solution and develop the Log Analyzer myself instead. I realized the analyzer in two parts: One part was done in Visual Basic (VBA) in Excel and the other part was done in Python.

The development of the Log Analyzer mainly consisted of two steps. The first step comprised a lot of manual cleaning, filtering and ordering of the log file (steps 1 to 5 below). In the steps 6, 7 and 8 below I developed the code for the Log Analyzer itself.

The steps that I took when building the analyzer (as outlined in Figure A.5 on page 110) were as follows:

1. I manually cleaned the log file of:

   (a) Advertising: Some browser events, like advertisements were logged by the Search-Logger along with information about the user behavior. I cleaned the log file from those ads (e.g. Google Adwords), by analyzing the log, identifying keywords (like "googleads" or "doubleclick", filtering for those keywords and filtering out those lines. By doing this I managed to reduce the log from from 30 000 to 20 000 log entries.

   (b) Automatically generated (e.g. by Javascript code in Web sites) non-user behavior related browser action such as logs related to the Facebook like button, Twitter button, Plusone button for Google's social network Google+. I used the same procedure as in (a)

   The time spent for this step totaled 2 weeks.

2. I translated the log entries into human readable format

   (a) I wrote an Excel macro to transform date and time entries into human readable format

   (b) I added a macro to translate number codes from questionnaires into human readable strings

   The effort for this step was about 1 week.

3. I analyzed how users changed queries during their searches. I first extracted the queries, which users entered into search engines by using the built in string functions in Excel. Then I analyzed those queries

   (a) regarding their length (number of words)

(b) regarding how a query was reformulated - looking at two sequential queries. I analyzed whether users broadened a query (omitting one or more search terms), narrowed a query (adding one or more search terms), entered a new query (no word in common with the previous query), used an equal query (equal search terms) and finally if they changed the query (substituted one or more search terms with other ones). For example if users entered the query "penicillin wiki" and then only "penicillin" this would be broadening. First using "penicillin wiki" and then "antibiotic" would be a new query. And first using "penicillin wiki" and then "penicillin blog" would be a query change.

(c) This step was quite significant as it took me quite a while to figure out a way to efficiently run this analysis with our rather large log file. The core of the analysis was as follows: We had e.g. the log entries user 1 "figur österreichisch kinderbuch", user 7 "Kasachstan", user 1 "Figur österreichischer Kinderbuchautor" and user 1 "Penicillin Erfinder". Queries 1,3 and 4 belonged to user 1, query 1 and 3 to one task and query 4 to another task. Query number 2 belonged to user 7. The sequence of queries needed to be analyzed according to the recipe explained in b). Of course in Excel it would have been straightforward to filter for users. Yet as I needed to take into account the timely sequence of the queries per users along with the task dependency, it made more sense to keep a certain number of log entries in the memory for this operation. I had the choice between Python and VBA to implement this procedure and opted for Python (as the common choice in the academic environment) to keep system specific dependencies low (Python ran on all machines in our research group).

The effort for this step was about 3 weeks.

4. I classified the log entries into several classes (from pages visited to SearchLogger specific events as outlined below). For example, to find all instances where users navigated to the second SERP in Google, I started with the "action description" column of the log file. I filtered out all entries that contained the word "Google". I looked for logs where the query reformulation was of type "equal" and where the logged search engine URL was identical. Another example is all user action related to image search. I filtered the log for the keyword "imgres?imgurl" in the action description. All URLs in the log that contained this string were instances where users clicked on an image in the Google image search engine. I

color coded the different classes in the log file to enhance readability. The classes were as follows:

(a) visited web sites

(b) search engine related logs; I analyzed the logs and identified the characteristics to classify them as follows:

   i. user visited a search engine (Google, Bing, Yahoo,..) and entered a query
   ii. user visited the an image search engine and issued a query
   iii. user visited the second search engine results page (SERP)
   iv. user entered a query into a news search engine

(c) usability related logs

   i. user opened a browser tab
   ii. user closed a browser tab

(d) Search-Logger specific events

   i. demographic form was displayed
   ii. user submitted demographics
   iii. user started a search case
   iv. user finished a search case
   v. user submitted a pre-task form
   vi. user submitted a post-task form

The effort for this step was about 5 weeks. The effort for this task was insofar significant as I had to analyze the logs first for their structure. Only after reverse engineering the URL structure of e.g. the Google news search engine, I could go on with carrying out the classification by implementing macros using the Excel string functions. This meant a lot of trial end error to get it right as each search engine uses its own URL structure and even within the Google family of search services a common format seems to be missing.

5. We manually analyzed the data that users submitted via demographic forms and pre- and post task forms. This step also included analyzing the search results users submitted and comparing them with the correct results.

   (a) demographic forms

i. We transformed the submitted demographic forms from the log into a separate Excel sheet. One line in the Excel sheet consisted of user number, age, gender, Internet usage per day (in hours), Internet usage per week (in days), usually used Web search engines and Internet user type specific information, e.g. if the Internet was more used for professional purposes or fun.

(b) pre- and post- task forms

i. We transformed pre- and post-task form data of the log file into a separate Excel sheet. This sheet contained 12 lines per user - 6 lines for the simple tasks and 6 lines for the complex tasks. Each line contained columns for the following user pre-task judgments: expected task complexity, expected time effort for task, expected query effort for task and if they expected to find the correct result. The post-task judgments were: experienced task complexity, experienced time effort for task, experienced query effort for task and if they thought they had found the correct result.

(c) search results

i. The results were submitted by users in the form of Word files. We manually analyzed the submitted results with the correct solution. We assigned to each solution a grade (1-4) whether it (1) was totally correct (2) partly correct (3) wrong or (4) no solution was submitted.

The effort for this step was about 4 weeks. I spent about 1 week on integration and analysis.

6. Finally I integrated all additional Excel sheets that contained data gathered during the experiment (demographics, pre-form, post-form) into one Excel file and made the data available to be used by the VBA analyzer:

   (a) First I extended the main data array to account for all the additional columns contained in the sheets mentioned above.

   (b) Then I changed all the corresponding procedures, which were used for reading all the data into the array.

   The effort for this step was 1 week.

7. As we started to gather ideas for publications I started adding different additional views at the data.

(a) For our publications we needed basically two views at the data

i. along tasks - to analyze the differences between simple and complex tasks. Therefore I needed to add a parameter to let the analyzer only take into account e.g. simple tasks or complex tasks.

ii. along users - to analyze the differences between individual people and also different user groups (e.g. good searchers vs. bad searchers). For this I added parameters to let the analyzer only take into account e.g. searches that led to correct results or searchers that led to incorrect results.

(b) I used Excel's filter functionality to make the data analyzable along any logged parameter: user number, user action, action description, time stamp, queries, reformulation type, query length and time for each step.

The effort for this step was about 3 weeks.

8. I analyzed the measures needed for our papers

(a) measures like (time per task, number of queries per task, number of queries per session, time per session, query length) were partly calculated in VBA or directly in Excel or I used a combined approach

(b) statistical analysis of mean values (2 sample t-tests, normal correlation coefficients, Spearman's rho correlation coefficients [90, 95]) was done in Excel's extended "Data analysis" environment

The effort for this step was about 4 weeks.

All steps that I took when building the analyzer module are illustrated in Figure A.5 on the next page. As can be seen in the flow chart, building the analyzer part included constant testing. In addition I ran one specific test after the log file was manually cleaned, translated into human readable format and the query reformulations were analyzed. I had to especially revise my algorithm to analyze the query reformulations several times till it was suitable for my requirements, e.g. accounting for instances where a query was unchanged. This could mean that the user entered the same query again but also that the user clicked on the next SERP. Then I implemented the functionality to classify the log entries and constantly tested the classification. Also this step included several iterations till the classification algorithm ran as desired. For example I continuously enhanced the classification by adding more types of user events to make the log file better understandable. Finally, I ran another set of tests

after the measures were extracted. I encountered several instances where I had to refine the whole analysis process to produce the measures in a format as wanted, e.g. when users switched tasks it meant that all time based measures had to account for that event.

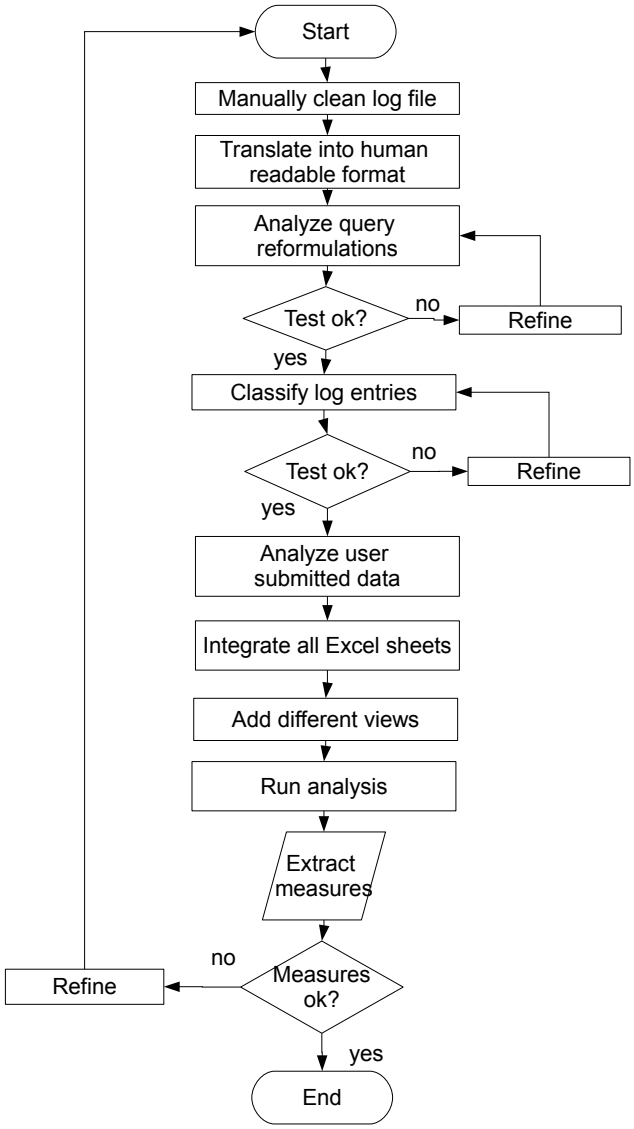

Figure A.5: Flow chart of analysis steps

In total the implementation of the analyzer part took me about 20 weeks (800 man hours).

Technical details and empirical benchmark data

I implemented the VBA based part in 3 VBA modules: Main, ReadIn and WriteOut.

The ReadIn module contains one method "Readoriginallogs" that reads the 8 data fields of the log file into the memory.

The Main module runs the calculations. It calls the ReadIn module (to read in the log data from the log files) and at the end of the calculation run it calls the WriteOut module to write out the results into a new Excel sheet.

The Main module consists of the following sub modules and calculates:

- SERP time per action (time on SERP per query)
- SERP per sequence time (time on SERP per search sequence without switching task=sum of all SERP times in previous step)
- SERP per task time (time on SERP per task=sum of all SERP sequence times)
- Reading time per action (time between two log events)
- Reading time per sequence (time spent for reading without switching the task)
- Reading time per task (sum of all reading times per sequence in previous step)
- Queries per task
- Type of queries (e.g. new, changed) per task
- Average query lengths
- Number of opened tabs and closed tabs per sequence and per task
- Number of opened tabs and closed tabs per task
- Reading to SERP ratio per task (Reading time/SERP time)

The sub-modules that caused a lot of trouble were the ones related to calculating SERP and reading time per task. It could happen that study participants started searching on a Task A for a while, which resulted in SERP and reading times for this search sequence. Then they switched to another Task B and started over with the previously started Task A at the end of the experiment.

The WriteOut module writes the final data out into an Excel sheet. The WriteOut module consists of the "user view" sub module (generating the data per user), "user summary view" module (generating the data for the whole sample averaging over users) and the "task summary view module" (generating the data for the whole sample averaging over tasks). As most results are mean values, they are written out together with their standard errors of mean.

This approach of using Excel and VBA was insofar instrumental as I always had good control over what was going on during the calculation runs, as WriteOuts were mainly in the same format as the ReadIns. Opting for VBA and Excel also had the advantage that the numbers were always visible and check-able.

After manually processing the log files and optimizing the algorithms for performance, one analysis run of 20 000 (cleaned) lines of log data takes less than a minute (significantly down from several minutes at the beginning). The results are written out in 5 additional Excel sheets showing the plane results of the analysis, the query change analysis, the user view, user summary view and task summary view.

Overall the log-file analysis was more challenging than initially expected. It comprised about 110 incremental iterations, done from the end of October 2011 till the beginning of February 2012. I have produced 1860 lines of code (LOC) in VB. The Main module consists of 797 LOC. The ReadIn module consists of 169 LOC. The WriteOut module consists of 894 LOC. The Python module (created together with Ulrich Norbisrath) consists of 120 LOC.

Problems and issues

During the process of analyzing the data I discovered mainly three areas that caused problems - the first was performance, the second was maintainability and the third was related to cooperation and sharing.

Performance Especially towards the end of the implementation phase, when most features had been implemented, one analysis run of the complete log file took 2-3 minutes depending on what variables I wanted to be analyzed. Iteratively adding changes to the code and testing became tedious. Therefore I created test log files (of significantly smaller size) containing logs of only e.g. 5-7 users and worked with those ones during time of code changes.

I also checked the code for unnecessary loops and if conditions and reduced the amount of write outs to a minimum in order to boost performance.

Maintainability At the very end of the analysis procedure when the complexity became gradually higher I started questioning whether Excel and VBA

(with its lack of support for object oriented programming features) were the appropriate tools to carry out this analysis with. Adding new functionality and parameters became increasingly time consuming.

Cooperation and sharing As Excel is a Microsoft specific product, I could not seamlessly share my code with e.g. my supervisor, who uses Linux. Sharing code e.g. via a joint repository would have been easier if the analyzer had been implemented in a platform independent programing language.

Excel was good for getting results quickly and learning the requirements and what to analyze. Therefore, the combination of Visual Basic and Excel was a good choice. Now that I know, what is needed (platform independent language, easier maintainability), I can start working on adding more functionality to the Python version and moving to a shared code model.

Appendix B

Publications

Abstract in Estonian

Veebi otsingumootorid ja vajadus keeruka informatsiooni järele

Veebi otsingumootorid on muutunud põhiliseks teabe hankimise vahenditeks internetist. Koos otsingumootorite kasvava populaarsusega on nende kasutusala kasvanud lihtsailt päringuilt vajaduseni küllaltki keeruka informatsiooni otsingu järele. Samas on ka akadeemiline huvi otsingu vastu hakanud liikuma lihtpäringute analüüsilt märksa keerukamate tegevuste suunas, mis hõlmavad ka pikemaid ajaraame. Praegused otsinguvahendid ei toeta selliseid tegevusi niivõrd hästi nagu lihtpäringute juhtu. Eriti kehtib see toe osas koondada mitme päringu tulemusi kokku sünteesides erinevate lihtotsingute tulemusi ühte uude dokumenti. Selline lähenemine on alles algfaasis ja ning motiveerib uurijaid arendama vastavaid vahendeid toetamaks taolisi informatsiooniotsingu ülesandeid.

Käesolevas dissertatsioonis esitatakse rida uurimistulemusi eesmärgiga muuta keeruliste otsingute tuge paremaks kasutades tänapäevaseid otsingumootoreid. Alameesmärkideks olid:

(a) arendada välja keeruliste otsingute mudel,

(b) mõõdikute loomine kompleksotsingute mudelile,

(c) eristada kompleksotsingu ülesandeid lihtotsingutest ning teha kindlaks, kas neid on võimalik mõõta leides ühtlasi lihtsaid mõõdikuid kirjeldamaks nende keerukust,

(d) analüüsida, kui erinevalt kasutajad käituvad sooritades keerukaid otsinguülesandeid kasutades veebi otsingumootoreid,

(e) uurida korrelatsiooni inimeste tava-veebikasutustavade ja nende otsingutulemuslikkuse vahel,

(f) kuidas inimestel läheb eelhinnates otsinguülesande raskusastet ja vajaminevat jõupingutust ning

(g) milline on soo ja vanuse mõju otsingu tulemuslikkusele.

Keeruka veebiotsingu ülesanded jaotatakse edukalt kolmeastmeliseks protsessiks. Esitatakse sellise protsessi mudel; seda protsessi on ühtlasi võimalik ka mõõta. Edasi näidatakse kompleksotsingu loomupäraseid omadusi, mis teevad selle eristatavaks lihtsamatest juhtudest ning näidatakse ära katsemeetod sooritamaks kompleksotsingu kasutaja-uuringuid. Demonstreeritakse põhilisi samme raamistiku "Search-Logger" (eelmainitud metodoloogia tehnilise teostuse) rakendamisel kasutaja-uuringutes. Esitatakse sellisel viisil teostatud uuringute tulemused. Lõpuks esitatakse ATMS meetodi realisatsioon ja rakendamine parandamaks kompleksotsingu vajaduste tuge kaasaegsetes otsingumootorites.

Käesolev dissertatsioon põhineb seitsmel autori artiklil, mis on publitseeritud teadusajakirjades ja konverentsikogumikes (või on kirjutamise hetkel trükis või hindamisel retsensentide poolt).

- Artikkel 1: Complex search: Aggregation, Discovery, and Synthesis [98]
  - Kirjandusülevaade, mis analüüsib olemasolevaid keerukate otsinguprotsesside kirjeldamise mudeleid .
  - Pakub välja mudeli, mis toetub selgelt defineeritutele ja mõõdetavatele agregeerimise, avastamise ja sünteesi kontseptsioonidele.
- Artikkel 2: Search-Logger - Analyzing Exploratory Search Tasks [102]
  - Esitleb vahendit kasutaja käitumise hindamiseks keerukate otsinguülesannete lahendamise käigus.
  - Esitab vastava pilootuuringu tulemused.
- Artikkel 3: Ordinary Search Engine Users Carrying Out Complex Search Tasks [101][esitatud hindamisele]
  - Tõestab, et keerukal otsingul on teatud erilised karakteristikud, mida on võimalik mõõta.
  - Esitleb mõõdikud, mis eristavad keeruka otsingu ülesanded lihtsaist ülesandeist.
  - Esitleb erinevaid karakteristikuid, mis eristavaid õigesti ja valesti sooritatud otsinguülesandeid.

  - Esitab proovitulemusi hästi- ja halvastisooritavate otsijate erinevuse kohta.

- Artikkel 4: Search Strategies of Library Search Experts [104]
  - Esitleb otsingustrateegiaid, mida kasutavad raamatukogu otsingueksperdid.
  - Esitab analüüsi internetiotsijate tüüpide mõju osas otsingutulemuslikkusele.
  - Toob sisse kaasajastatud klassifikatsiooni veebiotsingute strateegiate kohta.

- Artikkel 5: The Relationship between Internet User Type and User Performance when Carrying Out Simple vs. Complex Search Tasks [103]
  - Esitatakse interneti kasutajatüübi ja veebiotsingu tulemuslikkuse vahelist korrelatsiooni lihtsate ülesaanete juhul.
  - Esitatakse interneti kasutajatüübi ja veebiotsingu tulemuslikkuse vaheline korrelatsioon keerukate ülesaanete juhul.
  - Leitakse tulemuslikkuse kasutajatüübi-spetsiifiline erinevus lihtsa ja keeruka ülesande vahel.

- Artikkel 6: Ordinary Search Engine Users assessing Difficulty, Effort, and Outcome for Simple and Complex Search Tasks [100]
  - Esitab tava-veebikasutajate teostust ülesandes, kus nad peavad eelhindama keeruliste otsinguülesannete raskusastet, vajaminevat jõupingutust ja tulemuslikkust kasutades veebi otsingumootoreid.

- Artikkel 7: Impact of Gender and Age on Performing Search Tasks Online [99]
  - Esitleb soo- ja vanuseerisusi lihtsate ja keeruliste otsinguülesannete sooritamisel.

Curriculum vitae

General

- Name: Georg Singer
- Born: 12.10.1974, Mühlbach am Hochkönig, Austrian citizen
- Address: Kuperjanovi 48-1, 50409, Tartu, Estonia
- Email: Georg.Singer[@]ut.ee

Education

- 2009-2012 - University of Tartu, Institute of Computer Science, PhD student
- 2002-2004 - Danube University Krems, Business School, Master of Business Administration (MBA)
- 1994-1999 - University of Vienna, Institute of Theoretical Physics, Master of Science (MSc) in physics
- 1986-1993 - Gymnasium St. Johann, secondary education

Languages

- German (native speaker), English (very good), Estonian (moderate)

Work experience

- Since 2012 - Co-Founder Nootri Online Nutrition
- Since 2010 - Consultant, self employed
- Since 2009 - University of Tartu, Institute of Computer Science, lecturer

- 2008-2010 - Co-Founder Luxury Investments
- 2007-2008 - Pacific International, business development
- 2005-2007 - Alcan Austria, member of the board, business development
- 2001-2004 - Co-Founder Itcon
- 2000-2001 - Vienna University of Economics, research fellow

Awards and fellowships

- Since 2009 - PhD fellowship, Estonian Doctoral School
- 2011 Tiger University scholarship receiver

Elulugu

Üldandmed

- Ees- ja perekonnanimi: Georg Singer
- Sünniaeg ja -koht, kodakondsus: 12.10.1974, Mühlbach am Hochkönig, Austria, austerlane
- Aadress: Kuperjanovi 48-1, 50409, Tartu, Eesti
- Email: Georg.Singer[@]ut.ee

Haridus

- 2009-2012 - Tartu Ülikool, Arvutiteaduse Instituut, doktorant
- 2002-2004 - Donau Ülikool Krems, Austria, Business School, Master of Business Administration (MBA)
- 1994-1999 - Viini Ülikool, Teoreetilise Füüsika Instituut, magister scientiarum (MSc) füüsika
- 1986-1993 - Gümnaasium St. Johann, keskharidus

Keelteoskus

- saksa keel (emakeel), inglise keel (väga hea), eesti keel (kesktase)

Töökogemus

- Alates 2012 - Kaas-asutaja Nootri Online Nutrition
- Alates 2010 - Konsultant, füüsilisest isikust ettevõtja
- Alates 2009 - Tartu Ülikool, Arvutiteaduse Instituut, lektor

- 2008-2010 - Kaas-asutaja Luxury Investments
- 2007-2008 - Pacific International, ettevõtluse arendamine
- 2005-2007 - Alcan Austria, member of the board, ettevõtluse arendamine
- 2001-2004 - Kaas-asutaja Itcon
- 2000-2001 - Viini Majandusülikool, teadlane

Saadud uurimistoetused ja stipendiumid

- Alates 2009 - Eesti Doktorantide Kooli liige
- 2011 EITSA Tiigri Ülikool doktorandi stipendium

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. **Eno Tõnisson.** Solving of expession manipulation exercises in computer algebra systems. Tartu, 2002, 92 p.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.

43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.

64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.
74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
75. **Nadežda Bazunova.** Differential calculus $d^3 = 0$ on binary and ternary associative algebras. Tartu 2011, 99 p.
76. **Natalja Lepik.** Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Tartu 2011, 133 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
80. **Marje Johanson.** $M(r, s)$-ideals of compact operators. Tartu 2012, 103 p.