Anne Aula and
Klaus Nordhausen

Modeling Successful Performance in Web Search

DEPARTMENT OF COMPUTER SCIENCES
UNIVERSITY OF TAMPERE

D-2005-4

TAMPERE 2005

UNIVERSITY OF TAMPERE
DEPARTMENT OF COMPUTER SCIENCES
SERIES OF PUBLICATIONS D – NET PUBLICATIONS
D-2005-4, JUNE 2005

Anne Aula and Klaus Nordhausen

Modeling Successful Performance in Web Search

DEPARTMENT OF COMPUTER SCIENCES
FIN-33014 UNIVERSITY OF TAMPERE

ISBN 951-44-6355-2
ISSN 1795-4274

Modeling Successful Performance in Web Search

Anne Aula (corresponding author)

Tampere Unit for Computer Human Interaction

Department of Computer Sciences

FIN-33014 University of Tampere, Finland

Tel. +358-3-3551 8871, fax: +358-3-3551 6070

E-mail: anne.aula@cs.uta.fi

Klaus Nordhausen

Tampere School of Public Health

FIN-33014 University of Tampere, Finland

Tel. +358-3-3551 7086, fax: +358-3-3551 6057

E-mail: klaus.nordhausen@uta.fi

ABSTRACT

Several previous studies have measured differences in information search success of novices and experts. However, the definitions of novices and experts have varied greatly between the studies, and so have the measures used for search success. Instead of dividing the searchers into different groups based on their expertise, we chose to model search success with *Task Completion Speed*, TCS. Towards this goal, 22 participants performed three fact-finding tasks and two broader tasks in an observational user study. In our model, there were two variables related to the Web experience of the participants. Other variables included, for example, the speed of query iteration, the length of the queries, the proportion of precise queries, and the speed of evaluating result documents. Our results showed that the variables related to Web experience had expected effects on TCS. The increase in the years of Web use was related to improvement in TCS in the broader tasks, whereas the less frequent Web use was related to decrease in TCS in the fact-finding tasks. Other variables having significant effects on TCS in either of the task types were the speed of composing queries, the average number of query terms per query, the proportion of precise queries, and the participants' own evaluation of their search skills. In addition to the statistical models, we present several qualitative findings of the participants' search strategies. These results give valuable insight into the successful strategies in Web search beyond the previous knowledge of the expert - novice differences.

1. INTRODUCTION

Accessing information from the World Wide Web (the Web) has become routine behavior for a large user population. Different methods for accessing the information exist, such as browsing towards the information from a known page or using search engines. The browsing approach may be successful when the Web is used simply for finding something interesting, but for locating specific information, search engines are the most feasible and commonly used approach (Fallows, 2005).

It is essential to understand the process of information searching in order to design tools for information access. Several studies have looked at search engine log files to find out how typical search engine users search the Web (for a comparison of nine major log file studies, see Jansen & Spink, in press). These studies have revealed that the typical Web users only use a couple of query terms per query, have short search sessions, typically only check one result page, and often make mistakes with the advanced query operators, should they use them at all. Several researchers have expressed their concern over the fact that with these simple search strategies and the current tools available, casual Web searchers may face difficulties in locating the information they need (Hölscher & Strube, 2000; Jenkins *et al*. 2003; Leroy *et al.*, 2003). However, log studies give little information on the effectiveness of search strategies and thus, other empirical research methods are needed to determine whether these common strategies are successful. If the simple strategies are enough, there should be no reason to use the more complicated ones.

The diversity of the Web user population is vast. There are people who search information from the Web as their profession, there is a growing group of casual Web users who need to use the Web more or less regularly for work or for free-time information, and there is also a large group of users who use the Web only rarely. Web search engines are ideally meant to serve all of these people. User studies focusing on information searching from the Web have usually simplified the diversity of the user population by focusing on two groups, novices and experts. Most of these studies have established that experts are more successful in carrying out search tasks than novices (Jenkins *et al*., 2003; Hölscher & Strube, 2000), but there are also studies that have found only small differences between the experts' and novices' search strategies and outcomes (Brand-Gruwel *et al*., 2005).

There is a large variation in how the terms *novice* and *expert* are defined: some studies propose that users with a little over one working week (50 hours) of experience are already experts (Lazonder *et al*, 2000), others call final year PhD students in Educational Technology experts (Brand-Gruwel *et al*., 2005), and yet others include Internet professionals in their group of experts (Hölscher & Strube, 2000). The label novice, on the other hand, has been assigned to psychology freshmen (Brand-Gruwel *et al*., 2005), to people with less than five years of computer experience and less than one year of Internet/Web experience (Jenkins *et al*., 2003), to users who browse less than five hours per week (Khan & Locatis, 1998), or psychology students with one year of Web experience (Navarro-Prieto *et al*., 1999). In addition to the terminology related to search experience, also the terminology used for explaining the outcome of the search process varies considerably between studies. Different measures for success are used, such as task completion time or number of results found, making it difficult to compare the results from different studies.

In this study, the aim was to study how the different search strategies and background experience of the users affect the outcome in search tasks. Thus, we did not divide the users into novices and experts, but instead, we treated the level of experience as one possible predictor in a multiple linear regression model of search success. To model the possible factors affecting search success, we first defined a measure for search success called *Task Completion Speed* (TCS). TCS combines two typical search success measures, namely the task completion time (efficiency) and the number of tasks completed (effectiveness). The measure will be explained in more detail later in the paper.

The paper is organized as follows: first, we will review studies related to the effects of expertise and different search strategies in Web searching. After that, the methods used in this study will be explained. In Section 5, the model of variables affecting TCS will be presented. The model is then discussed and conclusions are drawn.

2. RELATED RESEARCH

There are numerous user studies about the effects of search expertise on the use of traditional information retrieval (IR) systems (for example, Fenichel, 1981; Fields *et al*., 2004; Hsieh-

Yee, 1993; Iivonen, 1995; Sutcliffe *et al*., 2000). However, the results from these studies cannot be directly applied to Web searching as marked differences have been found in the searching characteristics of Web and traditional IR system users (Jansen & Pooch, 2001). We will therefore specifically review studies on Web searching.

2.1. Defining experts and novices

Several authors have defined the terms *expert* and *novice* and characterized the differences between these groups. Table 1 shows how the users were divided into these two groups in eight studies, as well as the main findings of these studies.

(Table 1 should be positioned around here.)

Most frequently, Web experience (either the years of experience or the frequency of usage) has been used in dividing the users into novices and experts. However, the requirements for the experience levels have differed considerably between the studies. Thus, the "experts" in one study could easily be "novices" in another. This ambiguity in defining experience has certain consequences. First of all, the results from different studies cannot be directly compared and hence, the studies do not add to the understanding of the characteristics of the different groups. This has also the consequence that contradictory results from different studies may not actually be contradictory, but the differences may rather be due to very different groups of users being compared.

In addition to the division based on computer or Web experience, other approaches to defining experts and novices have also been used. For example, the division can be made post-hoc, based on the user’s success in certain tasks. In Leroy *et al*. (2003), the participants were divided into two groups based on their success and these groups were called *high*- and *low-achievers*. One final approach is to define experts as the ones with good strategies and novices as ones with “suboptimal” or “naïve” strategies (Sutcliffe & Ennis 1998; Fields *et al*, 2004).

Likewise, in psychological literature the definitions of expertise vary. Matlin (2002, p. 495) defines expertise as “consistently superior performance on a set of tasks for a domain, achieved by a deliberate practice over a period of at least 10 years”. In Eysenck and Keane (2000, p. 531), the definition of expertise is: “the specific knowledge that an expert has about a particular domain; *e.g.*, that an engineer might have about bridges, or a software engineer might have about programming.” Thus, Matlin emphasizes *both experience and performance* in her definition, whereas the definition of Eyesenck and Keane emphasizes on the *level of knowledge*. Higher knowledge can naturally be expected to result in superior performance, but this is not explicitly stated in the definition given by Eyesenck and Keane. Unexpectedly, the term *novice* is not defined in either of the books. Matlin (2002) simply begins to use this term when reviewing studies where experts are compared to non-experts in various problem-solving tasks. However, when generalizing the approach taken in numerous studies, a *novice* on something equals to being *inexperienced* on it. The inexperience is related to both quantitative and qualitative differences in task performance, for example, in different cognitive and motor tasks, as compared to the performance of experts (*e.g.*, Ashby & Maddox, 1992; Beilock *et al*., 2002; Novick, 1988; Johnson & Mervis, 1997). However, it should be noted that there can also be differences in the level of performance between inexperienced (novice) users. Thus, novices can further be divided into *good* and *poor performers* (De Jong & Ferguson-Hessler, 1986).

Shanteau *et al*. (2002) present nine different ways of identifying an expert. The most important ways in relation to the current study are listed below. The list is augmented with explanations inside the parenthesis to relate this listing to the domain of Web search:

- experience (most commonly used in Web search studies, however, the experience *in what* differs),
- certification (*e.g.*, librarians are called experts and clients are called novices),
- internal consistency (the reliability of the behavior should be high; could well be measured also in search studies by comparing the behavior over several tasks),
- behavioral characteristics (*e.g.*, self-confidence, creativity, stress tolerance; could also be measured in search studies by observing the behavior or from verbal protocols), and
- discrimination ability (experts can perceive subtle differences; in the process of information search, experts can be expected to discriminate between successful and less successful queries, relevant and irrelevant documents etc.).

Shanteau *et al*. highlight that none of the above-mentioned measures is sufficient alone, but a combination of measures is always needed. Concerning the relationship between expertise and experience, they state that "There are many examples of professionals with considerable experience who never become experts. (…) Although there are undoubtedly instances where a positive relationship exists between experience and performance, there is little reason to expect this to apply universally. At best, experience is an uncertain predictor of degree of expertise. At worst, experience reflects seniority – and little more." If we believe this statement to be true, the earlier studies using experience as the only measure for expertise are in a great risk of mislabeling individuals as experts, when they, in fact, are just experienced.

In sum, most of the Web search studies have defined experts and novices based on their computer/Web use/searching experience alone. However the need for other, more reliable, basis for this grouping is evident, when one takes into account the fact that experience does not necessarily make one an expert. One tempting approach is to use the level of performance in the actual search tasks for post-hoc dividing the users into experts and novices: experts could be defined as the ones whose performance was above a certain threshold, and novices as the ones whose performance was below the threshold. Towards this end, different ways of measuring the level of performance in search tasks are presented in the following section.

In addition to the expertise in the use of the tools needed for information search (computers, the Web, search engines), the expertise in the domain of the search is also known to affect search behavior. People who are familiar with the domain of the search can be expected to use high-quality search terms along with appropriate synonyms; they have even been claimed to be more systematic searchers who plan their searches beforehand (Hölscher & Strube, 2001; Jenkins *et al*., 2003; Navarro-Prieto *et al*., 1999). However, to benefit from the high domain expertise seems to require some experience with the search system (Hsieh-Yee, 1993; Vakkari *et al*., 2003). Generally, the searchers who can rely on both high domain and tool expertise complete search tasks most fluently (Hölscher & Strube, 2001; Jenkins *et al*., 2003).

2.2. Measuring the level of performance in search tasks

The two most important measures for the level of performance in searching are the task completion time (Brand-Gruwel *et al*., 2005; Jenkins *et al*., 2003; Khan & Locatis, 1998; Lazonder *et al*., 2000; Saito & Miwa, 2002) and the outcome of the search (task completion rate) (Hölscher & Strube, 2001; Jenkins *et al*., 2003; Khan & Locatis, 1998; Lazonder *et al*., 2000; Saito & Miwa, 2002). The outcome can be, for example, the number of relevant pages found or simply the correctness of an answer. If there are several tasks to solve, the total search outcome can be measured as the number of tasks solved successfully.

The obvious shortcoming in using task completion rate alone as a measure of success is that one searcher can be very slow but eventually find the correct answer, whereas another could find the answer almost immediately. We believe that people strive for efficient and effective performance also in information search tasks (along with the lines of the information foraging theory by Pirolli & Card, 1999), so there is a need for a measure that would define the faster user as the more successful searcher. However, if a searcher completed the task quickly but the quality of the answer was much poorer than that of someone who was a bit slower, we might be inclined to accept that the slower person was more successful. Thus, both effectiveness and efficiency are important in determining the success of search, but they need to be considered together, not independently as in most of the previous studies.

In addition to the task completion time and outcome, other measures for the efficiency of the search were also used in the studies listed in Table 1. For example, Palmquist and Kim (2000) measured the average time to retrieve a piece of information (a relevant page). Lazonder *et al*. (2000) measured efficiency as the ratio of successfully completed tasks to the time to complete these tasks. These measures have the benefit of combining information about the search success and search time. However, they do not take into account the time the user spent in unsuccessful tasks. This, in turn, leads to discarding all the data from participants who did not have any successful tasks. In addition, the efficiency estimated for participants with only a couple of successful tasks may be misleading (further discussion can be found in

Section 4.2). Efficiency measures related to browsing were also used, such as the number of pages viewed (Saito and Miwa, 2002) and the number of links traversed (Khan & Locatis, 1998). In our view, these efficiency measures do not always correlate with search success. We have observed that searchers sometimes benefit from browsing multiple documents as the documents help them to understand the topic of the search, and importantly, provide the searchers with additional search terms, if the initial query was not optimal.

In relation to the level of performance, experts typically outperformed novices at least in some measures used in the studies. Experts were faster (Saito & Miwa, 2002), more efficient (Palmquist & Kim, 2000), they had a better overall performance score (Khan & Locatis, 1998), or they outperformed novices in all the measures of the study (Lazonder *et al.*, 2000). However, there are also studies in which the experience did not affect the outcome of the search tasks (Brand-Gruwel *et al.*, 2005), no differences were reported (Navarro-Prieto *et al.*, 1999), or the differences between the groups were not statistically verified (Hölscher & Strube, 2001; Jenkins *et al.*, 2003). Two studies report speed differences (Brand-Gruwel *et al.*, 2005; Palmquist & Kim, 2000), but their statistical significance is marginal.

2.3. Characterizing the Strategies of Novices and Experts

Some studies have presented differences in the information search strategies of more and less experienced users, but the relation between these strategies and the level of performance is uncertain. The *queries* of experienced users tend to be longer than those of the less experienced users (Aula, 2003; Hölscher and Strube, 2000), *the use of Boolean operators and modifiers* appears to be higher among "professional" users (Aula & Siirtola, in press; Hölscher & Strube, 2003), and the experienced users are found to *commonly iterate their queries* when the information is not immediately found (Aula & Käki, 2003). Novices (or typical search engine users), on the other hand, frequently make errors in formulating sophisticated queries (Aula, 2003; Jansen *et al*., 2000). Navarro-Prieto *et al*. (1999) found that in fact-finding tasks, *the style of the queries* differed between more and less experienced users: experienced users frequently used a bottom-up strategy (the query terms are taken directly from the instructions), whereas the novices used very general queries (similar findings reported in Aula, 2003). Fields *et al*. (2004) reported that experts (librarians) showed excellent skills in reformulating their queries so that each reformulation built on the previous query and the retrieved results. However, they could not articulate these strategies when interviewed about them. On the other hand, non-experts (clients) issued a set of queries, each reflecting a *change in the approach* for finding the information. In relation to search success, it has been suggested that longer queries improve the precision of the result set and the search performance on the whole (Belkin *et al.*, 2003). However, precision might not be the top priority in Web search; Eastman (1999) even suggested that the use of advanced operators may not be worth the trouble in Web search engines. One study reported that the more successful searchers were already faster in *evaluating the results* given by the search engines meaning that they either iterated the query or selected a result quickly (Aula *et al*., in press). Although differences in search strategies have been found, the speed of performance of novices and experts has been found to be similar when the task is to find an answer to a specific task from the content of a certain Web page (Lazonder *et al*., 2000).

In sum, there are several studies on both the differences between experts' and novices' search strategies, as well as studies focusing on the performance differences in search tasks. However, studies focusing on both of these aspects of the search process at the same time are scarce. To fill this gap, our study aims to establish the relationship between different expertise-related strategies and search success. The variables are explained in detail in Section 4.1.

2.4. The approach taken in the current study

To avoid the problems related to the expert-novice division, we did not divide the users a priori into separate groups based on their experience or some other measure. Instead, we focused on the differences that the participants showed in their search strategies and the effects of these strategies in the participants' performance in search tasks (as measured by our novel measure, Task Completion Speed or TCS). Thus, we first ran the user studies and analyzed the participants' performance in search tasks. Following this, we analyzed which were the variables that resulted in high level performance, *i.e.*, which were the most successful or "expert strategies".

3. METHODS

3.1. Participants

The participants were recruited via advertisements in the notice boards of local universities and colleges. 22 people participated in the study (18 females and 4 males). The average age of the participants was 24.4 years (from 19 to 37 years, SD = 5.1). Of the participants, 20 were students of the University of Tampere and various local colleges, and two were researchers at the University of Tampere. The majors of the students varied, 4 majored in Computer Science, 4 in Translation studies, 6 in Psychology, and the rest in various different majors. All but one used computers at least *almost daily* with an average of 7.4 years of computer experience (from 2 to 15 years, SD = 4.8). 12 participants used the Web *daily*, and 10 less often than that. They had used the Web for 5.1 years on average (ranging from 1.5 to 10 years, SD = 2.6 years). In search engine usage, the frequency of use was not as high, with 10 participants using search engines at least *almost daily*, five people *several times a week*, and seven participants *weekly* or *less than weekly*. The median of the participants' own evaluation of their search skills was four on a five point scale (five was the best). Only one participant majoring in Information Science had formally studied information retrieval, whereas others were self-taught searchers. 21 participants mentioned using Google as their primary search engine, two mentioned using AltaVista in addition to Google, and one mentioned only using a Finnish search engine (http://www.fi).

3.2. Apparatus

During the search tasks, the participants used a PC workstation and a 19 inch monitor with a resolution of 1280 × 1024. The Internet connection was high-speed. All but one participant used Internet Explorer as the Web browser, the exception being Opera. During the search tasks, 21 participants used Google as the search engine (two of them used AskJeeves or AltaVista in addition to Google), and one participant used only a Finnish search engine, www.fi.

For the data analysis, the search sessions (both the screen image and the participants' facial expressions) were videotaped. Detailed log files of the videos were created with The Observer® version 5.0 software by Noldus information technology.

3.3. Tasks

In about 60% of the queries submitted to a search engine, the goal of the query is informational; the user's intent is to acquire information about the topic (Rose & Levinson, 2004). To study Web search strategies in relation to this most common task type, we prepared five search tasks with an informational goal (some of the tasks were taken or modified from previous studies (Aula, 2003; Aula, 2004)). To make sure that the difficulty for the tasks was appropriate for the current purpose, the tasks were pilot-tested with two participants. The topics for the tasks were chosen to be representative of the general topics that people use search engines for (for example, health, recreation, and computers) (Spink *et al.*, 2002). Three of the tasks were fact-finding tasks; where the goal was to find one specific answer to a simple question (tasks 1, 2, and 3) and two were broader, requiring the participant to find possibly several documents that would provide enough information for the task (tasks 4 and 5). Although tasks 4 and 5 were broader, they were still of closed format as the the participants could easily tell when they had found information that is relevant for the task. The tasks are listed below, with a short name for each task in parenthesis (this name will be used when referring to the tasks in the Results section):

1. How much blood does human heart pump in one minute? (Heart)
2. What is the most common dog breed in Finland as measured by the number of registrations? (Dog)
3. At what age do children typically learn to walk? (Children)
4. Find information on different antivirus software and their prices. (Virus)
5. Find information on how much energy do different foods contain and how much energy do different sports consume. (Energy)

In all of the tasks, the participants themselves decided on when they had found enough information for the task or when they wanted to give up on it.

In fact-finding tasks (*Heart*, *Dog*, and *Children*), the task was rated as *completed* if the participant found any document that contained the correct answer to the question. If the correct answer was not found, the task was rated as *not completed*.

In broader tasks, a three-level rating for task-completion was used. In all of these tasks, if the participants did not find any information concerning the task, the task was rated *not completed*. In task *Virus*, the task was rated as *completed* if the participant found at least two examples of antivirus software along with their prices. If they only found examples of software, but not the prices, the task was rated as *partially completed*. Similarly in task *Energy*, if the participant found at least two examples of different foods along with their energy contents, and at least two examples of sports along with the energy consumption, the task was rated as *completed*. If only sports or foods were covered in the answer, the task was rated as *partially completed*. In practice, both in tasks *Virus* and *Energy*, many participants found a list containing a lot more than two examples of the topics in question. In the cases where they searched for individual examples, they either themselves thought of two examples being enough or were told by the researcher that this information is enough for this task.

In sum, the fact-finding tasks were evaluated with two rating levels (completed or not completed) and the broader tasks were evaluated with three rating levels (completed, partially completed, or not completed).

3.4. Procedure

The studies were conducted in the usability laboratory at the University of Tampere. After arriving to the laboratory, the researcher informed the participants about the purpose and the procedure of the study. The participants were told that the purpose was “to study their normal information search strategies”. Thus, they were asked to complete the search tasks using the methods and tools (applications) that they would normally use. As the aim was to make the search session as typical as possible, it was also emphasized that the users could use search engines if they wished, but that they could also choose any other suitable methods – again, they should use the strategies they would normally use.

There was no strict time limit for the individual tasks. However, the time limit for the whole search session was limited to one hour. In practice, this meant that the participant could use as much time for the tasks as needed (at least some tasks were always quickly completed so that the one-hour time limit was seldom met). In only one case the researcher had to stop the task because of time constraints.

To collect the material they found, the participants were instructed to bookmark the pages they found relevant for the task. If the participant was not familiar with bookmarking, they were first shown how to make bookmarks with their browser of choice. The participants were told that they should evaluate the material themselves – whenever they felt that they had found enough information for the task, they could stop the task and move on to the next one. They were also instructed that if a certain task felt too difficult or for other reasons, they could stop the task at any time.

The participants were asked to think-aloud during the search tasks and explained what think-aloud means. In this explanation, we used similar instructions that are suggested by Ericson and Simon (1993). Following the instructions, the participants practiced both the procedure of the study and thinking aloud with one simple drawing task ("Draw a red circle with Paint"). The practice task also made it possible for the researcher to check that the procedure was understood correctly.

Before each task, the participants were given a piece of paper that contained the task description and two five-point scales. The participants evaluated their familiarity with the task (from very low to very high) and the difficulty of the task (from very difficult to very easy) using the scales.

At the end of the experiment, the participants were briefly interviewed about the experiences during the search session and asked to fill in a questionnaire about their background information (*e.g.*, age, computer and Web experience). The questionnaire asked for the following background information:

- years of computer experience,
- years of Web experience,
- frequency of using search engines (options: daily, almost daily, couple of times a week, weekly, less than weekly),
- frequency of using computers (options: daily, almost daily, couple of times a week, weekly, less than weekly),
- frequency of using the Web (options: daily, almost daily, couple of times a week, weekly, less than weekly), and
- own evaluation of search skills on a scale from 1 (unskilled) to 5 (skilled).

Finally, the participants were thanked and given a movie ticket as a compensation for their time.

4. DATA ANALYSIS

4.1. Variables Included in the Model

The modeling approach selected for the current study requires the variables in the model to be independent of each other. As many of the experience-related variables are not independent (for example, one cannot use the Web or search engines more often than computers), we needed to choose only the most important experience-related variables to the models. Web experience is typically used as the basis for dividing participants into experts and novices (for example, in Hölscher & Strube, 2001; Jenkins *et al.*, 2003; Khan & Locatis, 1998; Lazonder *et al.*, 2000). Thus, we also chose Web experience over computer and search engine experience to include in the models. The experience-related variables the models contain are “years of Web experience” and “frequency of using the Web”. As the participants’ own evaluation of their search skills is not logically dependent on their Web experience, we also included this variable in the models.

In the question concerning the frequency of using the Web, some of the options only contained a couple of responses. Thus, the options of this question were combined so that the variable had two levels (daily and less than daily). In the participants’ evaluation of their own search skills, the options were also re-grouped so that there were two levels for this variable, *skilled* (ratings 4 and 5) and *unskilled* (ratings 1-3).

From the video transcripts, we formulated the following variables:

- *Proportion of precise queries of all queries*. Precise query was defined as a query that covers all of the main aspects (facets) of the task (definition adopted from Aula, 2003). This variable was chosen to the model as earlier research (Aula, 2003; Navarro-Prieto *et al.*, 1999) has suggested that experienced and less experienced users formulate different kinds of queries (broad and precise)
- *Number of queries per minute.* This measure was calculated by dividing the number of queries per task with the total time the user spent on the search engine’s pages (query and result listing) per task. Thus, the time the user spent inspecting the result

documents did not affect this measure. This variable was included in the model as Aula *et al.* (in press) suggested that more experienced users tend to evaluate the results faster than less experienced users before iterating the query or selecting a result document.

- *Average number of query terms per query.* A term was defined "as a string of characters separated by some delimiter such as a space, a colon, or a period" (Jansen and Pooch, 2001). Queries of experienced users have found to be longer than those of less experienced users (Aula, 2003; Hölscher & Strube, 2000), which was the reason for including this variable.
- *Percentage of the task time spent inspecting result documents.* This is the proportion of the total task time the user spent in result documents (other pages than the search engine). When searching for a specific piece of information from a certain Web page, the performance of less and more experienced users should not be different (Lazonder *et al*., 2000). This variable was included to the model to test whether this result holds when the Web pages need to be inspected when performing a search task with a search engine.
- *Average number of results opened per query.* This is the average number of documents the user selected from the search engine's result listing for further inspection. Shanteau *et al*. (2002) defined "discrimination ability" as one criterion of expertise. In the case of web searching, the more successful searchers are expected to show a greater discrimination ability by having a smaller number of results opened per query. In tasks where the answer can be found from one Web page, the value of this variable should be close to one.

This list is not a comprehensive list of strategies that are related to expertise. However, due to the modeling approach requiring independent variables, we could not include all of the variables that have been suggested in earlier studies (*e.g.*, “number of changes in querying approach” measure would be related to the “proportion of precise queries, and thus, it could not be included in the model).

In addition to the variables described above, we calculated the number of times each participant used different advanced search strategies, such as the Web browser’s Find command to locate certain strings from the document, Boolean operators, and term modifiers. These measures were not included as variables in the model, as their usage was quite infrequent. The queries were also analyzed qualitatively in order to find the possible difficulties in query formulation. Furthermore, we transcribed and analyzed all the verbalizations the users made during the test.

The task difficulty and familiarity measures were discarded from the statistical models. This was done because the models were formulated by averaging the measures over several tasks. Average evaluation of familiarity or difficulty would not have been informative in the models as the averages from different people were close to each other.

4.2. Measure for Task Completion Speed, TCS

In information retrieval studies, search success is often measured by *precision* (the proportion of relevant material among the retrieved results) and *recall* (the proportion of relevant material retrieved from the document collection) (Baeza-Yates & Ribeiro-Neto, 1999, p. 75). Being system oriented and requiring a well-known search environment, these measures are not suitable for the current purposes. To account for the dynamic nature of user-system interaction, *interactive recall* (ratio of truly relevant documents that were judged relevant by the user) and *interactive precision* (ratio of documents evaluated as relevant by the user that were truly relevant) measures have been proposed by Veerasamy & Heikes (1997). However, they are also inappropriate for the goals of the current study. First, in search tasks where the goal is to answer a specific question, recall is not relevant: if the answer to the question is found from one page, the other pages providing the same answer will not provide any added benefit for the searcher. The measure of *precision* would not provide much information in the current case either: nearly all of the pages the users bookmarked were relevant, so the precision would in most cases be 1. Thus, this measure would not be informative in distinguishing between less and more successful searchers.

As noted earlier, task completion rate and task completion speed are commonly used to measure the performance differences between novices and experts. However, we feel that it is essential to consider these variables together to avoid the risk of considering fast but unsuccessful performance as superior to slow but successful. In addition, we argue that fast performance should be considered superior to slow, if the task completion level is the same. Thus, to measure the performance in the search tasks, two variables, task time and task completion, were combined together to form a single measure, *Task Completion Speed* (TCS). In order to attain a high TCS, the searcher's behavior needs to be *consistent* (fast and accurate) in all tasks, which was presented as one of the characteristics of expertise by Shanteau *et al*. (2002). The TCS measure was calculated with the following formula:

$$TCS = \frac{No\_of\_tasks\_completed}{Total\_search\_time\_\sec} \times 3600$$

In this formula, the quotient (number of tasks completed/task completion time for all tasks) is multiplied by 3600 in order to make the speed of the magnitude tasks/hour. The task completion was evaluated from the contents of the pages the participants bookmarked during the test (the rating scheme was explained earlier in Section 3.3).

The TCS measure is similar to the *efficiency* measure by Lazonder *et al*. (2000). However, in their measure, the time used in the formula contains only the time for the successfully completed tasks. In contrast, we also included the time used for the unsuccessful tasks. There are a couple of reasons for doing this. First, if using only successfully completed tasks in the analysis, important data is lost. Data from some participants would be discarded altogether, if they did not have any successfully completed tasks. Second, we believe that by taking into account the time for unsuccessful tasks, we will get a better estimate of the participant's overall search speed. An example will clarify the issue: if the participant only completed one task (out of three) successfully during the test and she used 1 minute for this task (let's assume that her average time for the unsuccessful tasks was 10 minutes), her *efficiency* score would be 1.0. Another user, who completed all three tasks successfully and used 3 minutes for each, would also have an efficiency score of 1.0. Thus, both of these users would be judged as equally efficient, which intuitively, does not seem to be the case. The same example is used below to compare the users with our TCS measure.

In order to use the task completion ratings in the TCS measure, they were transformed into numbers (not completed = 0, partially completed = 0.5, completed = 1). These ratings can be thought of as the proportion of task completion: 0 means that the task was not completed at all, 0.5 means that half of the task was completed, and 1 means that the task was fully completed. These ratings were then summed up over all tasks. If the participant got 1 "completed", 1 "partially completed" and 1 "not completed" ratings, his/her total number of tasks completed would be 1.5 (1+0.5+0). This number is then divided by the total time the participant used for searching in all tasks (completed, partially completed, and not completed). Thus, larger values of TCS imply better search performance. If considering how TCS would judge the two hypothetical users presented above, the first user's TCS would be (1 successful task/1260sec) × 3600 = 2.86 and the second user's TSC would be (3 successful tasks/540 sec) × 3600 = 20. Thus, the second user's performance would be judged as better.

4.3. Modeling TCS

We modeled TCS using a multiple linear regression model (Venables & Ripley, 2002, Chapter 6). This model explains TCS as a linear function of the variables described above, namely the five experience-related variables, the participants' own evaluation of their search skills, and the seven variables related to search strategies. For an optimal model, we used a backward selection method based on the Akaike Information Criterion (AIC). This criterion is defined as

AIC = -2 × loglikelihood of the model + 2 × the number of parameters in the model.

The backward selection begins with the full model and removes variables step-by-step until the model with the lowest AIC is found. AIC was chosen because we had a high number of variables in our model as compared to the number of participants (AIC punishes the number of variables in the model).

To evaluate whether quadratic terms of variables should be included in the model, we used a visual inspection of scatter plots. For the final model, we checked the standard model diagnostics (*i.e.,* several residual plots and Cook's distance) to evaluate the model assumptions. To identify possible influential subjects, we related the Cook's distance to the F distribution as described in Neter *et al.* (1996, 381). The analysis was performed using the R 2.0.1 statistical software package (R development Core Team, 2004).

4.4. Analysis of the Qualitative Data

The purpose of the qualitative data analysis was to gain more information about the behavioral characteristics of the searchers by observing their behavior and by the use of the think-aloud protocol. In the analysis of the verbalizations (think-aloud) during the search tasks, the verbalizations were first transcribed. Following this, the transcriptions were printed out. These print-outs were then read through several times while at the same time marking important events in them. Following this, the events were grouped into categories that emerged from the data.

5. RESULTS

Descriptive statistics of the variables per task (averages and ranges per variable) are presented in Table 2. Table 3 shows the task completion times, task completion rates, as well as the Task Completion Speed (TCS) per participant in both fact-finding and broader tasks. These tables show that the task times were an average of 95 seconds faster in fact-finding tasks than in the broader tasks; in fact-finding tasks, the tasks were also successfully completed more often than in broader tasks (86.4% vs. 79.5% of the tasks, respectively). The number of query terms per query was found to be typical for Web searches (average over all tasks was 2.22). In the fact-finding tasks, the queries tended to contain more terms (2.56, on average) than in the broader tasks (1.86, on average).

(Table 2 should be positioned around here.)

(Table 3 should be positioned around here.)

In this section, we will first present the model for the broader tasks (*broader model*), followed by the model for the fact-finding tasks (*fact-finding model)*. In the models, the variables were added as quadratic terms if the scatterplots indicated a quadratic, rather than linear relationship. The variables in quadratic form were interpreted by assuming an optimal maximum or minimum point, assuming that all the other variables are held fixed. Moving away from this point in either direction increases or decreases the TCS level depending on the sign of the estimate for the quadratic term. The optimal point is often negative, although in practice, the variable can only have positive values. To make the interpretation of the effects of the variables easier in these cases, we will set the optimal point to 0 and only discuss the values with practical meaning (the negative values will not be discussed).

Since we used the AIC selection criterion and not a criterion based on p-values, the final models contain also non-significant variables. In the following, we will concentrate on the significant variables (Sig. < .05). However, also the variables that are marginally significant (Sig. < .10) will be discussed.

5.1. Model for the Broader Tasks

The model for the broader task included data from two tasks, namely tasks *Virus* and *Energy*. In this model, Participant 19 appeared to be an influential observation (Cook's distance 0.98, corresponds to the 49.41 percentile of the F(10, 12) distribution). However, in the scatterplots, the data of this participant did not stand out as deviant from the data of other participants. Thus, we decided to retain the data in the model. The full model with 8 original variables (plus three variables in a quadratic form) is presented in Appendix A (Table A1). Table 4 presents the final model after backwards selection.

(Table 4 should be positioned around here.)

In this model, the years of Web experience had a quadratic effect on TCS with an optimal minimum point at 0. Thus, people with longer Web experience outperformed those with less experience. Surprisingly, the participants' own evaluation of their search skills did not match their objective performance level with those who evaluated their search skills lower having a higher TCS.

In this model, an increase in the proportion of precise queries had a negative effect on TCS. Thus, as the proportion of precise queries increased, TCS decreased. Along the same lines, the quadratic variable measuring the average number of query terms was marginally significant with an optimal maximum point at 0.38. As queries always have at least one query term, an increase in the number of query terms was related to a decrease in TCS. The reasons for these slightly counterintuitive findings will be discussed in Chapter 6.

5.2. Model for the Fact-Finding Tasks

The model for the fact-finding tasks included data from tasks *Heart*, *Dog*, and *Children*. In this model, participant six was found to be an influential subject in variable measuring the number of queries per minute (a Cook's distance of 1.43, which corresponds to the 72.63 percentile of the F(10, 12) distribution). A closer inspection of the scatterplots indicated that the value this participant had for the variable "average queries per minute" was highly unusual with the participant having a very high querying speed and a very low TCS. Thus, the data from this participant was discarded from the analysis. The full regression model with 11 variables (8 variables selected from the original variables presented above, plus 3 variables in a quadratic form) is given in Appendix A (Table A2). The final model (Table 5) has six of the original variables plus three variables in quadratic form.

(Table 5 should be positioned around here.)

In this model, the only significant variable is the average number of result documents opened per query (the more documents the user opens per query, the lower the TCS). Furthermore, the model fit in general is poor with $R^2 = 0.568$. When checking the model fit from the fitted and observed values, it was evident that the model significantly underestimated the TCS for Participant 13 (observed TCS = 23.46, fitted TCS = 14.38, residual = 9.08), whereas the residuals for other participants were much smaller (between -5.40 and 3.04) (the possible reasons for this are presented in Section 6). Thus, we decided to model the data again without Participant 13. This time, no variables were removed in the backwards selection. Thus, the full and final model is presented in Table 6 with 11 variables (8 variables selected from the original variables presented above, plus 3 variables in a quadratic form).

(Table 6 should be positioned around here.)

By discarding Participant 13, the model fit improved considerably (from adjusted $R^2 = 0.568$ to adjusted $R^2 = 0.797$). The significant variables of this model are presented next.

The variable measuring the frequency of the Web use affected TCS so that those using the Web less than daily had a lower TCS than the daily users. An unexpected finding was that the searchers who evaluated themselves as less skillful, seemed to perform better than those who thought of themselves as more skillful (significance marginal).

The variable measuring the speed of querying (in quadratic form) had an optimal minimum point at 0 meaning that TCS increases with an increase in the speed of querying. The variable measuring the average number of query terms also showed a quadratic effect on TCS with an optimal minimum point at 0. Thus, longer queries are related to increase in TCS.

5.3. Qualitative Analysis

The think-aloud data, observations, as well as the qualitative analysis of the queries provided several insights into the behavioral characteristics of the participants during the searching, as well as reasons for the problems they encountered. In the following, we will explain the main findings along with transcribed verbalizations from the participants as illustrations. To make it easier to relate the qualitative findings to the participants' performance in the search tasks, we included the TCS of the particular task type in parenthesis after the identification number of the participant.

None of the participants explicitly entered Boolean operators in their queries. The use of term modifiers was also rare. Phrase search (quotation marks around the terms) was the most frequent "advanced search style" used by 18.2 % of the participants. In addition, one participant excluded terms from the query by using Google's advanced search page. There were also two users who separated terms with the word 'and' which is excluded from the queries in Google (as a Boolean operator, 'and' needs to be capitalized and as it is the default operator in Google, it is not needed in the queries). Five users (27.3 % of the participants) used '+' to separate the query terms instead of space (the '+' sign is also excluded from the queries in Google). Thus, in most of the cases, the use of the advanced features was erroneous. These errors themselves did not affect the participants’ search success in Google, as they were simply omitted from the queries. However, the average TCS of the participants making these errors seemed to be low: in fact-finding tasks, their average TCS was 8.03 (the average of all participants was 12.34) and in broader tasks, their average TCS was 5.18 as compared with the average of 8.57 for all participants. 27 % of the participants used the Web browser’s Find command to quickly locate important terms from the result documents.

Intuitively, the query formulation might be expected to start with the cognitively simplest query formulation. Clearly defined tasks, such as the ones used in this study, often provide good query terms already in the task description and one would expect the easiest strategy to be to use the terms from the task description as query terms. Despite of this, participants often formulated queries where the terms were not directly taken from the task description but instead, the query terms were abstracted from the wording used in the tasks. Especially in fact-finding tasks, these generic and abstract queries were often associated with the user expressing feelings of uncertainty about the query formulation. If the participants used the simple approach of taking the query terms from the task description, they often did so after trying other approaches first. For example, Participant 20 (TCS = 6.85) first formulated the following queries (translated from Finnish) for the task *Dog*:

- dog tax
- dog tax Finland
- statistics
- registration dog
- number of registrations dog

Finally, after these trials, this participant formulated the successful query ‘most common dog breed’, and uttered: “How frustrating, if it turns out to be this simple…” In the same task, Participant 9 (TCS = 13.60) formulated the query ‘most common dog breed in Finland’ after four unsuccessful queries and uttered: “Could it be this easy? If I simply use the same terms that are in the task description…?” In task *Heart*, Participant 3 (TCS = 4.56) had a similar reaction after deciding to formulate the query ‘heart pumps blood per minute’ after two unsuccessful queries: “Ok, let’s try it like this. If the result is found with this query, I will be so disappointed that I did not formulate this query earlier.” In all, eleven participants commented similarly about it being almost “too simple” to directly use the terms from the task.

Based on our observations, the less successful searchers often choose the next query or action after an unsuccessful attempt so that it looks almost random; they do not seem to systematically modify their next query based on the results they receive. In the current case (task *Heart*), Participant 22 (TCS = 2.63) formulated the queries ‘human heart’, ‘circulation heart’, ‘circulation in the heart per minute’, ‘heart blood pumps’, ‘heart blood pumps liter’, and finally ‘heart blood minute’. In task *Energy*, Participant 9 (TCS = 5.34) formulated the queries ‘energy food’, ‘nutrition’, ‘how much energy contain’, ‘nutritional value’, and ‘energy consumption sports’ (plus others). On the other hand, successful searchers often systematically reformulated the queries. For example, Participant 19 (TCS = 14.85) formulated the following queries for the task *Children:*

- children learn to walk at age
- children learn to walk at age references
- typically children “learn to” walk age

- typically children “learn to walk” age
- typically “children learn to walk” age

As these queries were not successful, the participant changed the approach and formulated the query ‘developmental psychology learn to walk’ which provided the needed information.

Earlier research has suggested that Web searchers, especially less experienced ones, use short queries and seldom iterate them. The think-aloud data provided information about the reasons for using this possibly sub-optimal querying style. The following think-aloud examples make it clear that the formulation and iteration of queries is difficult:

Participant 3 (TCS = 4.71) in task *Energy*: “You need to have a rich imagination to think up the search terms!”

Participant 4 (TCS = 3.09) in task *Energy*: “Often, the terms are there but the results are not related to the topic. I really do not know how I could make the search terms to be relevant for this task.”

In addition to the explicit verbalizations, the observations showed that participants often spent considerable amounts of time to think about the query terms to use. Although the exact times used for this thinking were not analyzed, the observations suggested that the less successful searchers might use more time in this phase of the search.

6. DISCUSSION

The approach of the current study agrees with the ideas of Sutcliffe (1998), Fields *et al.* (2004), and Matlin (2002), who state that expertise implies superior strategies and performance. Furthermore, in line with Shanteau *et al.* (2002), we believe that experience does not necessarily make one an expert and therefore, experience alone should not be used to divide people into experts and novices. Instead of using experience as a basis for dividing people into different groups, we focused on the level of performance in actual search tasks as an indication of expertise. This enabled us to see whether we can find successful and efficient strategies that may in the end, be the strategies that experts use. Towards this goal, we included eight variables (plus three variables in a quadratic form) in multiple linear regression models of task completion speed (TCS) for fact-finding and broader search tasks. In the following, the final models (in fact-finding tasks, the model with 20 participants) are discussed in detail.

In the final model for the broader tasks, the Web experience as measured by the years of Web use had an expected effect with the more experience resulting in higher TCS. The frequency of using the Web was not included in the final model. In fact-finding tasks, on the other hand, the effect of the years of Web use was not strong enough to be included in the final model. In these tasks, the frequency of using the Web was related to TCS so that those using the Web less than daily had a lower TCS than those using it every day. These results support the notion that experience is related to expertise, although only one of the measures (frequency and years of use) was significant in each model. We propose that the possibilities of using a measure that combines the usage frequency and the years of Web experience should be studied in the future. In the present case, the usage frequency was collected with a questionnaire having pre-defined values (daily, almost daily, several times a week, weekly, less than weekly). This approach is not accurate enough to be used for the combined experience measure. Instead, the frequency of use should be measured as precisely as possible, for example, as the hours of Web use per week. The combined measure could be calculated by multiplying the frequency of using the Web (*e.g.*, hours per week) by the years of Web experience (we have already used a similar way of calculating computer experience in Aula *et al.*, in press). The product could further be multiplied by 52 to make the result to be interpreted as *total experience in hours*. For example, a user with 4 years of Web experience with an average of 10 hours of weekly use would get a score of 4 × 10 × 52 = 2080, while another user with 8 years of Web experience with an average of 2 hours of weekly use would get a score of 8 × 2 × 52 = 832. We believe that this would be a more reliable measure of the overall Web usage experience than the traditional experience measures alone.

The participants' own evaluation of their search skills was a poor predictor of the level of performance in fact-finding tasks as those evaluating themselves as unskilled seemed to perform better than their skilled counterparts. In the broader tasks, the relationship between the evaluated skill level and TCS was as expected with the unskilled participants having a lower TCS than the skilled participants. Based on these contradictory findings, we believe that the subjective evaluation of the search skills cannot be used reliably in defining experts and novices in information search tasks.

Earlier research suggests that searchers should express their information need to the search system as thoroughly as possible, implying that the longer and more precise the queries, the better the search success (Belkin *et al.*, 2003). The analysis of the *fact-finding tasks* showed that the longer queries resulted even in a quadratic improvement in TCS. However, the model of the TCS in *broader tasks* suggested that the longer queries decreased TCS (the effect of the variable was marginally significant), as did the increase in the proportion of precise queries of all queries. Thus, the claim that longer queries improve search performance does not seem to hold for all search task types. In broader tasks, it is often more difficult to think of the specific query to use in order to find the needed information. Because of this, it may be advantageous to locate documents that are related to the topic (with a shorter query) rather than try to guess the terms to use in the query. However, it should be noted that Belkin *et al.* (2003) probably expect the longer queries to be beneficial only if they are used optimally. In the case of the broader task *Energy*, a good query with six query terms might look like this: '(food OR nutrition) AND (sports OR exercise) AND (calories OR energy)'. In reality, a query with multiple terms is likely to be more along these lines: 'energy consumption sports food contents' (a real query by Participant 6). As mentioned earlier, sophisticated use of Boolean operators is rare in Web search. Thus, although the "more information in queries provides better results" claim may be true in theory, in the life of a typical Web searcher, this relationship may not hold.

In the fact-finding model, the speed of querying was retained in the final model as a significant variable showing that as the speed of querying increases, so does TCS. The speed of querying includes the time it takes for the user to formulate the query and evaluate the results before performing the next action (selecting a result or re-formulating the query). This result is in line with the findings by Aula *et al*. (in press) who showed that a fast evaluation of the search results is beneficial at least in certain circumstances.

This study showed again that the Web searchers' use of Boolean operators and term modifiers is rare and error-prone. Although their usage has earlier been related to search expertise (Hölscher & Strube, 2000), their usefulness in Web searching has been questioned altogether (Eastman & Jansen, 2003). Furthermore, studies have found their usage to be rare even among highly experienced users (Aula *et al*., 2005), the only exceptions being professional users (Aula & Siirtola, in press). The rare usage of the Boolean operators and term modifiers by the participants of the current study could be seen as indicating that there were no "real search experts” in our sample. However, we feel that the participants’ varying level of experience and presumably expertise, as well, was adequate for the current modeling purposes, were they any “real experts” (however they may be defined) among the participants or not (if expertise was defined as in the studies listed in Table 1, nearly all of our participants would be called experts). In addition, we believe that also the highly successful searchers (in our view, the real experts) optimize their strategies: if the simple strategies without operators and modifiers are enough, they will not employ the more complex ones just because they are able to do so.

The exploratory analysis of the queries and the think-aloud protocols suggested several variables whose relationship to search performance should be studied further. For example, participants were often reluctant to simply use the terms from the task description and instead, they chose to use more generic concepts as query terms (e.g., a query *human anatomy* for the task Heart). This tendency may indicate that participants conceptualize the searching from the Web as similar to using an encyclopedia to find information. If using an encyclopedia, a successful strategy is to think of the generic concept in relation to the information need whereas in Web search, this strategy often results in immensely large result sets.

The strategy of using terms that are irrelevant for the current task but that are presumably found from the relevant result page was used by some participants (this strategy was also revealed by Aula *et al*., 2005). The relationship between this strategy and search success needs to be addressed in further studies.

We believe that TCS is a feasible measure for search success (and expertise in searching) when the study setup consists of multiple tasks, when the level of task completion is relatively straightforward to rate (in this study, a three-level rating – completed, partially completed and not completed – was used for broader tasks, and a two-level rating – completed and not completed – for fact-finding tasks), and when the tasks are similar enough to be combined into a single measure. In some earlier studies, the users have only completed one task. We strongly discourage this approach: the level of performance in one task might not be representative of the searchers typical level of performance. With several tasks, the effect of coincidental fluctuations in the level of performance can be minimized. We also believe that the TCS measure could be utilized in different kinds of search tasks than the ones used in this study. For example, if the task was to find all the relevant documents in the database (and this number was known or could be estimated), it would be straightforward to calculate the task completion rate for each task by dividing the number of found documents by the number of all relevant documents in the database (similarly than in the standard measure of recall). Following this, the task completion rates could be added together for the "number of tasks completed" value needed in the TCS formula.

Oftentimes, when the so-called expert strategies are found, it is concluded that "novices" should be trained to use these strategies. However, we believe this training approach to be practically impossible: there are hundreds of millions of people using search engines regularly – and we believe that most of these people are not interested in searching per se (l*et al*one training in it). Thus, instead of claiming that people should be taught to use less or more query terms and precise or broad queries depending on the task, we believe that the search engines should simply make the use of these strategies possibilities for all users. For example, as the participants' verbalizations showed, it is very difficult for the user to think of additional or alternative query terms. Thus, search engines should provide the users with carefully chosen alternative terms. Fortunately, some of the publicly available search engines already do this.

This study presented a novel approach for studying expertise in Web search by modeling successful search behavior instead of dividing the users into two groups and measuring the differences in them. The study should be seen a starting point for understanding the relationship between different search strategies and TCS in different types of search tasks. Being the first step, the study left some open questions for further studies. For example, the analysis of the queries showed that mistakes in query formulation (unnecessary 'and' or '+' between the query terms), although dismissed by Google, seemed to be related to lower TCS. Also the other variables that could not be included in the model this time deserve attention in the future studies. The relatively small sample size (22 participants) and the small number of tasks per task type resulted in the models being mathematically weaker than would be the case with larger samples. Thus, the results should be confirmed and the exploration of the successful Web search strategies continued with also larger-scale studies.

The variables that were found to explain the differences in TCS (as well as variables that were identified in the qualitative analysis) could be useful in the quest of building *proactive search engines.* Based on the behavior of the user, these search engines could identify problematic situations and provide context-sensitive assistance for the user. Although there is still a long way to go before the search engine can determine whether the search strategies are successful or unsuccessful, we believe the modeling of search success to be a feasible approach for providing the necessary background information before proactive search engines become reality.

7. CONCLUSIONS

The approach taken in this study was that expertise in information search is manifested by the use of beneficial search strategies. Thus, we believe that through increased expertise in information search, the users select strategies that lead to more successful results. We defined success as efficient and effective behavior in search tasks (along the lines of the information foraging theory (Pirolli & Card, 1998)). To measure success along these lines, we introduced a measure called Task Completion Speed (TCS) and provided insights into the variables that affect TCS in search tasks with Web search engines. The variables measuring the level of Web experience (years and frequency of Web use) showed expected effects on TCS with the more experience resulting in increased TCS. We also discussed the possibilities of using a combined measure for Web experience that might provide benefits over the traditional experience measures. In addition to experience-related variables, our study showed that the level of performance can also be explained by certain behavioral variables, such as the speed of composing queries, the average number of query terms per query, and the proportion of precise queries.

Our approach showed that expertise can be treated as a continuous quality instead of a static state with two values (expert and novice). We feel that it is actually somewhat misguided to treat expertise as a dichotomous variable with two separate values; there are many shades of gray between the two extremes. In particular, by simply dividing people into two groups from a certain average value and then calling one group experts and the other group novices inevitably leads to internally heterogeneous groups. Naturally, an easy remedy to the arbitrary definitions of experts and novices is to replace the terms with *experienced* and *less experienced*. However, the question of *experienced in what* still remains.

8. ACKNOWLEDGMENTS

The authors would like to thank all of the participants for their time. Professor Ann Blandford, Tomi Heimonen, Esa Eteläperä, and Mika Käki, as well as anonymous reviewers are thanked for their valuable comments on the manuscript. This study was financially supported by the Graduate School in User-Centered Information Technology (UCIT) and the Academy of Finland (project 178099).

REFERENCES

Ashby, F.G. & Maddoz, W.T. (1992) Complex decision rules in categorization: contrasting novice and experienced performance. Journal of Experimental Psychology: Human Perception and Performance, 18, 1, 50-71.

Aula, A. (2003). Query formulation in information search. In Isaías, P. & Karmakar, N. (Eds.), Proceedings of the IADIS International Conference WWW/Internet 2003 (pp. 403-410). Lisboa: IADIS Press.

Aula, A. (2004). Enhancing the readability of search result summaries. In A. Dearden, & L. Watts (Eds.), Proceedings Volume 2 of the Conference HCI 2004: Design for Life, Leeds, UK, September 2004 (pp. 1-4).

Aula, A., Jhaveri, N., & Käki, M. (2005) Information search and re-access strategies of experienced web users. Proceedings of WWW 2005, May 10-14, 2005, 583-592, Chiba, Japan.

Aula, A., & Käki, M. (2003). Understanding Expert Search Strategies for Designing User-Friendly Search Interfaces. In P. Isaías, & N. Karmakar (Eds.), Proceedings of the IADIS International Conference WWW/Internet 2003 (pp. 759-762). Lisboa: IADIS Press.

Aula, A., Majaranta, P., & Räihä, K.-J. (2005). Eye-tracking reveals the individual strategies for search results evaluation. To appear in Proceedings of INTERACT’2005, Rome, Italy, September 2005.

Aula, A., & Siirtola, H. Hundreds of folders or one ugly pile – strategies for information search and re-access. To appear in Proceedings of INTERACT 2005, Rome, Italy, September 2005.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999) *Modern Information Retrieval*. New York: ACM Press.

Beilock, S.L., Carr, T.H., MacMahon, C., & Starkes, J.L. (2002) When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. Journal of Experimental Psychology: Applied, 8, 1, 6-16.

Belkin, N.J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., Tang, M.-C., & Yuan, X.-J. (2003). Query length in interactive information retrieval. Proceedings of the 26th Annual International Conference on Research and Development in Information Retrieval (SIGIR'03) (pp. 205-212). New York: ACM Press.

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novice: analysis of a complex cognitive skill. Computers in Human Behavior, 21, 487-508.

De Jong, T. & Ferguson-Hessler, M.G.M. (1986) Cognitive structures of good and poor novice problem solvers in physics. Journal of Educational Psychology, 78, 4, 279-288.

Eastman, C.M. (1999). 30,000 hits may be better than 300: precision anomalies in Internet searches. Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR'99) (pp. 313-314). New York: ACM Press.

Eastman, C.M. & Jansen, B.J. (2003) Coverage, relevance, and ranking: the impact of query operators on web search engine results. ACM Transactions on Information Systems, 21, 4, 383-411.

Ericsson, K.A. & Simon, H.A. (1993) Protocol analysis: Verbal reports as data. Revised edition. Cambridge, MA: The MIT Press.

Eyesenck, M.W. & Keane, M.T. (2000) Cognitive psychology: A student's handbook. Hove: Psychology Press.

Fallows, D. (2005) Search engine users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve. Pew Internet & American Life Project. Retrieved June 5, 2005, from http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf.

Fenichel, C.H. (1981). Online searching: Measures that discriminate among users with different types of experiences. Journal of American Society for Information Science, 1, 23-32.

Fields, B., Keith, S., & Blandford, A. (2004). Designing for expert information finding strategies. In S. Fincher, P. Markopoulos, D. Moore, & R. Ruddle (Eds.) People and Computers XVIII – Design for Life. Proceedings of HCI 2004 (pp. 89-102).

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science, 44,* 3, 161-174.

Hölscher, C., & Strube, G. (2000). Web search behavior of internet experts and newbies. Proceedings of 9th International WWW conference. Amsterdam, The Netherlands (pp. 337-346).

Iivonen, M. (1995). Searchers and searchers: differences between the most and least consistent searchers. Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR’95) (pp. 149-157).

Jansen, B. J., & Pooch, U. (2001). Web user studies: A review and framework for future work. Journal of the American Society for Information Science and Technology. 52, 3, 235 - 246.

Jansen, B.J. & Spink, A. (in press) How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing and Management.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. Information Processing and Management. 36, 2, 207-227.

Jenkins, C., Corritore, C.L., & Wiedenbeck, S. (2003). Pattens of information seeking on the Web: A qualitative study of domain expertise and Web expertise. IT & Society, 1, 3, 64-89.

Johnson, K.E. & Mervis, C.B. (1997) Effects of varying levels of expertise on the basic level of categorization. Journal of Experimental Psychology: General, 126, 3, 248-277.

Khan, K. & Locatis, C. (1998). Searching through the cyberspace: the effects of link display and link density on information retrieval from hypertext on the World Wide Web. Journal of the American Society for Information Science, 49, 2, 176-182.

Lazonder, A.W., Biemans, H.J.A., & Worpeis, I.G.J.H. (2000). Differences between novice and experienced users in searching information on the World Wide Web. Journal of the American Society for Information Science, 51, 6, 576-581.

Leroy, G., Lally, A.M., & Chen, H. (2003). The use of dynamic contexts to improve casual Internet searching. ACM Transactions on Information Systems, 21, 3, 229-253.

Matlin, M. (2002). Cognition. New York: Thomson Learning.

Navarro-Prieto, R., Scaife, M., & Rogers, Y. (1999). Cognitive strategies in Web searching. Proceedings of the 5th Conference on Human Factors & the Web. Retrieved June 5, 2005, http://zing.ncsl.nist.gov/hfWeb/proceedings/navarro-prieto/.

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. (1996). Applied Linear Statistical Models, 4 th edition. Chicago: Irwin.

Novick, L.R. (1988) Analogical transfer, problem similarity, and expertise. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 3, 510-520.

Palmquist, R.A., & Kim, K-S. (2000). Cognitive style and on-line database search experience as predictors of Web search performance. Journal of the American Society for Information Science, 51, 6, 558-566.

Pirolli, P. & Card, S. (1998). Information foraging. Psychological review, 106, 4, 643-675.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL: http://www.R-project.org. ISBN 3-900051-00-3.

Rose, D.E. & Levinson, D. (2004) Understanding user goals in Web search. Proceedings of WWW 2004, 13-19.

Saito, H., & Miwa, K. (2002). A cognitive study of information seeking process in the WWW: the effects of searcher’s knowledge and experience. Proceedings of the Second

International Conference on Web Information Systems Engineering (WISE’01) (pp. 0321). IEEE.

Shanteau, J., Weiss, D.J., Thomas, R.P., & Pounds, J.C. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. European Journal of Operational Research, 136, 253-263.

Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002) From e-sex to e-commerce: Web search changes. IEEE Computer, 55, 3, 107-109.

Sutcliffe, A., & Ennis, M. (1998). Towards a cognitive theory of information retrieval. Interacting with computers, 10, 321-351.

Sutcliffe, A.G., Ennis, M., & Watkinson, S.J. (2000). Empirical studies on end-user information searching. Journal of the American Society for Information Science, 51, 13, 1211-1231.

Vakkari, P., Pennanen, M., & Serola, S. (2003) Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing and Management*, 39, 3, 445-463.

Venables, W.N., & Ripley, B.D. (2002). Modern Applied Statistics with S, 4th edition. New York: Springer-Verlag.

APPENDIX A: FULL MODELS

(Table A1 should be positioned around here).
(Table A2 should be positioned around here).

Table 1: User studies focusing on the differences in information search strategies between less and more experienced users.

Study	Experts	Novices	Measures for success	Main differences in strategies or success
Brand-Gruwel et al. (2005)	Final year PhD students in the field of Educational Technology	Psychology freshmen	Task time and the quality of a 400-word argument.	Time difference marginally significant (experts spent more time in the task). No difference in the quality of the assignment.
Hölscher & Strube (2001)	Internet professionals with at least 3 years of intensive experience and a daily use of the Internet as a source of information	No explanation (novices identified by an interview and a pre-test, which are not described)	Rate of solving the tasks.	Double experts (domain & Web expert) solved more tasks than other groups (no statistical analysis done). Double novices (domain & Web novices) make more query re-formulations. Web experts use query formatting tools more often and make less errors than novices.
Jenkins et al. (2003)	> 5 years of computer use. 4.5 years of Internet, Web and search engine use.	1 - 5 years of computer use, less than 1 year of Internet/Web use.	Search time and outcome.	Double-experts faster than the other groups (no statistical analysis done). Web novices tend to use a breadth-first strategy whereas experts use a depth-first strategy.
Khan & Locatis (1998)	More than 5 hours of browsing per week.	Less than 5 hours of browsing per week.	Overall performance, accuracy, number of links traversed, search time & prioritization.	Experts had a better overall performance score, which was mostly due to them prioritizing tasks better.
Lazonder et al. (2000)	Over 50 hours of www experience, self-reported proficiency ranged from 8 to 12.	Less than 10 hours of Web experience, self-reported proficiency < 5.	Time, success, efficiency, effectiveness.	Experts outperformed the novices in all of the measures.
Navarro-Prieto et al. (1999)	(*Experienced Web users*) Computer science students, 2 years of experience.	Psychology students, one year of Web experience.	None.	Experienced users choose different strategies (bottom-up, tow-down, or mixed) according to the task they are conducting.
Palmquist & Kim (2000) – tasks carried out on a Web site, no search engines.	Participants divided on “experienced” and novice users based on a questionnaire asking for how long and frequently the participant has been using online databases and if they could name the databases they are using.		Average time, efficiency (total number of nodes visited/the number of bookmarks).	Experienced users faster and more efficient than novices.
Saito & Miwa (2002)	Participants divided on experts and novice users based on a questionnaire. The questionnaire had questions about daily Web use, information seeking style on the Web, and knowledge about search engines.		Result (relevant target found), time, number of pages browsed, kinds of pages browsed	Experts were faster and referred to fewer pages than novices.

Table 2: The average values of the variables per task. The range (minimum, maximum) is presented in parenthesis.

	Fact-finding tasks			Broader tasks	
Variable	**Heart**	**Dog**	**Children**	**Virus**	**Energy**
Task time (seconds)	264.44 (25.48, 734.60)	294.63 (42.24, 553.08)	327.23 (82.24, 705.72)	327.13 (85.60, 574.60)	454.37 (132.00, 878.92)
Task completion rating	0.82 (0.00, 1.00)	1.00 (1.00, 1.00)	0.77 (0.00, 1.00)	0.82 (0.50, 1.00)	0.77 (0.00, 1.00)
Proportion of precise queries (%)	51.70 (0.00, 100.00)	69.44 (14.29, 100.00)	60.77 (0.00, 100.00)	30.68 (0.00, 100.00)	57.83 (0.00, 100.00)
Number of queries per minute	2.11 (0.85, 6.15)	1.89 (0.28, 3.96)	1.64 (0.00, 4.34)	2.13 (0.00, 7.56)	2.16 (0.00, 5.76)
Number of query terms per query	2.62 (1.33, 7.00)	2.30 (1.50, 3.50)	2.77 (2.00, 5.17)	1.79 (0.00, 3.00)	1.98 (0.00, 3.50)
Percentage of task time spent inspecting result documents	53.46 (0.00, 88.67)	48.45 (13.35, 79.35)	51.98 (13.05, 85.11)	68.05 (36.70, 91.33)	63.05 (39.30, 92.20)
Number of result documents opened per query	3.00 (0.00, 9.00)	3.91 (1.00, 10.00)	4.41 (1.00, 10.00)	4.05 (1.00, 11.00)	5.09 (1.00, 14.00)

Table 3: The percentage of tasks completed and the average task completion time by participant in fact-finding and broader tasks. TCS refers to the average Task Completion Speed in these tasks.

	Fact-finding tasks			Broader tasks		
Participant	% of tasks completed	Average task time	TCS	% of tasks completed	Average task time	TCS
1	100.0	455.7	7.9	75.0	434.3	6.2
2	100.0	214.0	16.8	100.0	361.9	10.0
3	66.7	526.1	4.6	50.0	382.2	4.7
4	66.7	273.0	8.8	50.0	581.8	3.1
5	100.0	286.3	12.6	100.0	375.0	9.6
6	33.3	182.8	6.6	75.0	572.8	4.7
7	100.0	165.3	21.8	100.0	473.3	7.6
8	100.0	399.1	9.0	75.0	403.1	6.7
9	100.0	264.7	13.6	75.0	505.2	5.3
10	100.0	364.2	9.9	50.0	394.3	4.6
11	66.7	425.7	5.6	100.0	408.6	8.8
12	100.0	302.3	11.9	50.0	543.8	3.3
13	100.0	153.4	23.5	75.0	554.8	4.9
14	100.0	169.3	21.3	75.0	436.2	6.2
15	100.0	168.1	21.4	50.0	365.4	4.9
16	66.7	278.6	8.6	100.0	240.4	15.0
17	100.0	191.8	18.8	100.0	274.1	13.1
18	100.0	233.9	15.4	100.0	329.6	10.9
19	100.0	242.4	14.9	100.0	179.7	20.0
20	66.7	350.3	6.9	75.0	143.5	18.8
21	100.0	397.1	9.1	75.0	337.7	8.0
22	33.3	455.5	2.6	100.0	298.9	12.0
Average	86.4	295.4	12.3	79.5	390.7	8.6

Table 4: Final model for TCS in broader tasks (n = 22).

Significance of the model $p < .001$ $R^2 = .889$ (adjusted $R^2 = .802$) AIC = 39.677	Estimate	Standard error	t	Sig.
Intercept	17.759	1.950	9.106	<.001
Years of Web experience (linear)	5.551	2.489	2.230	.046
Years of Web experience (quadratic)	6.924	3.294	2.102	.057
Frequency of using the Web (less than daily)	1.667	1.128	1.478	.165
Own evaluation of search skill (novice)	-3.610	1.083	-3.334	.006
Proportion of precise queries	-0.087	0.030	-2.918	.013
Average number of query terms (linear)	3.678	3.232	1.138	.277
Average number of query terms (quadratic)	-4.894	2.617	-1.870	.086
Percentage of task time spent inspecting result documents	-0.011	0.008	-1.382	.192
Average number of result documents opened per query	-0.396	0.260	-1.520	.154

Table 5: Full model for TCS in fact-finding tasks (Participant 6 excluded, n = 21).

Significance of the model p = .018 R^2 = .762 (adjusted R^2 = .568) AIC = 65.086	Estimate	Standard error	t	Sig.
Intercept	22.010	3.985	5.524	< 0.001
Years of Web experience (linear)	-9.900	5.694	-1.739	.110
Years of Web experience (quadratic)	-10.617	6.479	-1.639	.130
Frequency of using the Web (less than daily)	-4.220	2.537	-1.664	.124
Own evaluation of search skill (novice)	4.690	2.989	1.569	.145
Number of queries per minute (linear)	7.881	4.531	1.739	.110
Number of queries per minute (quadratic)	-2.144	6.517	-0.329	.748
Average number of query terms per query (linear)	5.498	5.410	1.016	.331
Average number of query terms per query (quadratic)	13.681	7.822	1.749	.108
Average number of result documents opened per query	-2.558	0.989	-2.587	.025

Table 6: Final model for TCS in fact-finding tasks (Participants 6 and 13 excluded, n = 20).

Significance of the model p = .004	Estimate	Standard	t	Sig.
R^2 = .915 (adjusted R^2 = .797)		error		
AIC = 43.847				
Intercept	26.963	5.399	4.994	.001
Years of Web experience (linear)	-7.464	4.429	-1.685	.130
Years of Web experience (quadratic)	-7.646	4.350	-1.758	.117
Frequency of using the Web (less than daily)	-5.251	1.695	-3.097	.015
Own evaluation of search skill (novice)	4.464	1.962	2.276	.052
Proportion of precise queries	-0.077	0.058	-1.329	.221
Number of queries per minute (linear)	11.250	3.691	3.048	.016
Number of queries per minute (quadratic)	1.727	5.210	0.331	.749
Average number of query terms per query (linear)	8.319	6.765	1.230	.254
Average number of query terms per query (quadratic)	13.766	5.587	2.464	.039
Percentage of task time spent inspecting result documents	-0.083	0.087	-0.948	.371
Average number of result documents opened per query	-1.556	0.853	-1.825	.105

Table A1: Full model for TCS in broader tasks (n = 22).

Significance of the model p = .002 R^2 = .888 (adjusted R^2 = .764) AIC = 43.547	Estimate	Standard error	t	Sig.
Intercept	17.734	2.163	8.198	<.001
Years of Web experience (linear)	5.578	2.748	2.030	.070
Years of Web experience (quadratic)	7.017	3.631	1.933	.082
Frequency of using the Web (less than daily)	1.759	1.291	1.363	.203
Own evaluation of search skill (novice)	-3.548	1.302	-2.724	.021
Proportion of precise queries	-0.089	0.038	-2.360	.040
Number of queries per minute (linear)	-0.477	2.996	-0.159	.877
Number of queries per minute (quadratic)	-0.455	2.802	-0.162	.874
Average number of query terms per query (linear)	3.617	3.757	0.963	.358
Average number of query terms per query (quadratic)	-5.088	2.996	-1.698	.120
Percentage of task time spent inspecting result documents	-0.011	0.009	-1.237	.244
Average number of result documents opened per query	-0.392	0.288	-1.364	.203

Table A2: Full model for TCS in fact-finding tasks (Participant 6 excluded, n = 21).

Significance of the model p = .064 R^2 = .776 (adjusted R^2 = .503) AIC = 67.801	Estimate	Standard error	t	Sig.
Intercept	20.124	8.586	2.344	.044
Years of Web experience (linear)	-7.046	7.387	-0.954	.365
Years of Web experience (quadratic)	-12.090	7.317	-1.652	.133
Frequency of using the Web (less than daily)	-4.782	2.823	-1.694	.125
Own evaluation of search skill (novice)	5.029	3.266	1.540	.158
Proportion of precise queries	0.046	0.082	0.560	.589
Number of queries per minute (linear)	9.847	6.098	1.615	.141
Number of queries per minute (quadratic)	-1.446	8.775	-0.165	.873
Average number of query terms per query (linear)	-1.133	10.622	-0.107	.917
Average number of query terms per query (quadratic)	16.604	9.331	1.779	.109
Percentage of task time spent inspecting result documents	-0.052	0.145	-0.361	.727
Average number of result documents opened per query	-2.070	1.408	-1.470	.176