Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 82–91

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Controlling the complexity in comparing search user interfaces via user studies

Mika Käki [a,*], Anne Aula [b]

[a] Idean Enterprises Ltd., Kauppakatu 3 A 11, FI-33100 Tampere, Finland
[b] Google, Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94042, United States

Received 10 April 2006; received in revised form 31 January 2007; accepted 19 February 2007
Available online 6 April 2007

Abstract

Over time, researchers have acknowledged the importance of understanding the users' strategies in the design of search systems. However, when involving users in the comparison of search systems, methodological challenges still exist as researchers are pondering on how to handle the variability that human participants bring to the comparisons. This paper present methods for controlling the complexity of user-centered evaluations of search user interfaces through within-subjects designs, balanced task sets, time limitations, pre-formulated queries, cached result pages, and through limiting the users' access to result documents. Additionally, we will present our experiences in using three measures – search speed, qualified search speed, and immediate accuracy – to facilitate the comparison of different search systems over studies.

Keywords: Search user interfaces; User-centered evaluation; Methodology; Evaluation measures

1. Introduction

User-centered studies on information search are becoming increasingly common mostly due to the increasing popularity of Web searching. Traditionally, information retrieval (IR) studies have been system-centered, focusing on the performance of the algorithms matching queries with relevant documents. However, interest in the users' search strategies has increased in the IR community over the years, a notable example of this interest being the interactive track in the TREC conference.

In the field of human computer interaction (HCI), the focus is by definition on the human side of computing systems. Thus, in HCI studies, the involvement of users in the evaluation and comparison of computer systems has a long tradition. Although the importance of including the users in the comparisons of search systems is acknowledged both in the field of HCI and IR, combining the system and user-centered approach has proven to be surprisingly difficult; researchers conducting user-centered studies on search user interfaces have had trouble in finding suitable forums to publish their work.

* Corresponding author. Tel.: +358 44 5353298.
E-mail addresses: mika.kaki@gmail.com (M. Käki), anneaula@google.com (A. Aula).

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.02.006

In Tampere Unit for Computer–Human Interaction (TAUCHI) at the University of Tampere, search user interfaces have been studied extensively with a user-centered approach over the past years. This paper shares our experiences on the methodological issues we have learned when conducting experimental comparisons of Web search user interfaces. This paper provides a coherent view of the methods and measures we used in our studies. Individual studies along with methods and metrics have previously been presented in separate publications, but this paper focuses on the strengths and weaknesses of these methods and metrics in order to assist researchers to choose the appropriate methodological approach when planning and conducting experimental comparisons of search user interfaces.

We will discuss two issues that play a major role in user-centered search user interface comparisons:

1. Variance brought to the test setup by the involvement of users. This variance can compromise the validity of the experiment if not handled properly as the measured effects may be smaller than the variance caused by the user's personal characteristics. Based on our experience, relatively radical actions can and should be used to control the variance present in the user-centered search studies.
2. The use of appropriate measures. Appropriate measures enable and facilitate comparisons between studies and allow robust and useful conclusions to be made.

Partial solutions to these problems can be found from the previous studies reporting individual experiments. This paper combines the methodological suggestions and discusses their implications in detail.

Of the four phases of searching (query formulation, search execution, result evaluation, and possible query reformulation), our studies have mainly focused on enhancing the result evaluation phase through new user interface solutions. In evaluating these solutions, we have successfully employed balanced task sets, limited the users' time in the search tasks, used pre-formulated queries, and restricted the users' access to result documents. Certainly, the restrictions raise questions about ecological validity, but they can be largely overcome by utilizing the methods when the research questions are suitable for an experimental study and by conducting complementary studies with other methods, such as log studies (Käki, 2005b).

In measuring the success of search user interfaces, we have employed the existing measures of interactive precision and recall (Veerasamy & Belkin, 1996; Veerasamy & Heikes, 1997). However, in the course of our research, a need for measures that would capture the characteristic user behavior in web searching rose. Thus, we proposed two search speed measures and a measure called *immediate accuracy* to capture the success specifically in Web searching. These measures have revealed new insights on the systems and are a good addition to the toolbox for user-center evaluations.

2. Related work

Although the evaluation of search systems becomes more challenging when users are involved in the process, the popularity of this approach has grown over the years (Savage-Knepshield & Belkin, 1999). There are numerous methods for collecting data about the use of search systems, such as transaction logs (Jansen & Pooch, 2001; Silverstein, Marais, Henzinger, & Moricz, 1999; Spink, Wolfram, Jansen, & Saracevic, 2001), observing the use of a search system (Aula & Käki, 2003; Aula & Käki, 2005; Jones, Bruce, & Dumais, 2001), questionnaires (Aula, 2003; Aula, Jhaveri, & Käki, 2005), and experimental studies (Brajnik, Mizzaro, & Tasso, 1996; Dumais, Cutrell, & Chen, 2001; Hertzum & Frøkjær, 1996; Heimonen & Jhaveri, 2005; Paek, Dumais, & Logan, 2004; White, Ruthven, & Jose, 2002).

In user-centered comparisons of search user interfaces, the comparison is often made in relation to the three main aspects of usability, namely, effectiveness, efficiency and satisfaction, as defined in ISO-9241-1 standard (1998). In this standard, effectiveness is defined as the "accuracy and completeness with which the users achieve specific goals" with the system, efficiency refers to the needed resources for achieving goals with the system, and satisfaction to the user's attitudes towards using the system. In user-centered evaluations of search interfaces, it is important to measure all of these aspects of usability as the different aspects do not always correlate (Frøkjær, Hertzum, & Hornbæk, 2000).

The results of the search interface comparisons are often reported in figures related to effectiveness (e.g., percentage of tasks completed) and/or efficiency (e.g., task times) (White et al., 2002; Woodruff, Faulring, Rosenholtz, Morrison, & Pirolli, 2001). Often, these two figures are reported separately. In addition to these objective measures for search success, the role of the user's perceived success and subjective satisfaction has been emphasized by several researchers (Su, 1994; Tague & Schultz, 1988; White, Ruthven, & Jose, 2003). The most common methods for collecting information on subjective satisfaction are interviews (White et al., 2002; Woodruff et al., 2001) and questionnaires (Brajnik et al., 1996; White et al., 2002, 2003).

When comparing the usability of search user interfaces, the study setup needs to be controlled to assure that the observed differences are due to the differences in the interfaces rather than some confounding variables. However, the level of control varies between studies: users can perform tasks in their normal context while the researcher remotely logs their actions (Anick, 2003; Hertzum & Frøkjær, 1996) or the study can be conducted in a laboratory where the researcher can carefully control the situation (Topi & Lucas, 2005; Woodruff et al., 2001). In the former, the ecological validity of the study is perhaps higher, but the results may otherwise be harder to interpret due to the uncontrolled factors.

We have concentrated in experimental studied that have mostly been carried out in a laboratory setting. Such an approach allows and also requires a great deal of control. Next, we will discuss the controlling methods we have employed.

3. Controlling complexity

Controlling the large number of variables that enter the test situation along with the user is a major challenge in user-centered studies. We have successfully utilized several methods to ensure that only the phenomenon of interest, namely, the user interface, varies while the other factors are controlled.

3.1. Experimental design

The users' characteristics such as search and domain expertise, age, and cognitive style are known to have a large effect on the search process (Aula, 2003; Aula & Käki, 2003; Aula et al., 2005; Ford, Miller, & Moss, 2005; Lazonder, Biemans, & Worpeis, 2000). In addition, the type of the search task has an effect on the strategies that people use (White & Iivonen, 2001).

To control the effect of the user-related factors, we find the *within-subjects design* to be appropriate. In this design, all the participants use all the tested interfaces and thus, the users' personal search strategies remain constant over the tested interfaces and variation decreases. This conclusion is in line with several TREC interactive track guidelines from the past years (e.g., TREC interactive track guidelines, 1999, 2000, 2002).

Although being an attractive solution by minimizing the variability caused by individual differences, the within-subject design poses also challenges. One problem is that the presentation order of the compared systems may have an effect on the users' performance and preferences. This, however, can normally be controlled by counterbalancing the presentation order between the users.

Another and a more serious shortcoming in the within-subjects design is that the same tasks cannot be used with each compared interface because the participants learn the solutions and successful strategies on the first encounter. The obvious solution of using different tasks may, however, introduce additional variance to the setup.

To decrease the variance caused by the different tasks, we have used *balanced task sets*. To use balanced task sets, equally many task sets are required than there are conditions in the test. Each task set is used once in a test, in one condition. To make the task sets as similar as possible, we have found it useful to create pairs (or number of interfaces to be compared) of tasks, the difficulty and topics of which are as similar as possible. In practice, it is often enough to just modify one term from the task description, for example: "Find Chinese restaurants in New York" and "Find Chinese restaurants in Los Angeles". Keeping the topics in the pairs of tasks constant is advisable as it reduces the effects that the users' interest in different topics may have on their performance.

For creating realistic tasks, we have used common topics from actual Web searches (e.g., Spink, Jansen, Wolfram, & Saracevic, 2002). Although we have verified the successfulness of the task set balancing with pilot tests, in reality, it is practically impossible to create two tasks sets that would be equally demanding. Thus, it is always advisable to statistically analyze whether the task sets had an effect on the dependent variables of the study despite of all the precautions.

Another problem with balanced task sets, in addition to them being laborious to construct, is that they may introduce some learning effects. In the example above, it is possible that the user discovers a smart search strategy for that kind of a problem on the first encounter. The second condition would then benefit from the learning effect. We have simply used the test moderator's judgment to decrease this risk by removing the tasks affected by learning effects based on the pilot test observations.

For controlling the variation caused by different task sets, we have used a *counterbalanced* design so that each task set is used with each interface equally many times and that the presentation order of the task sets and interfaces is balanced. For such balancing, the well-known Latin Square is a useful aid. While counterbalancing removes major issues, it is not a magic tool to make test setup right. Selection of the tasks and balancing of the task sets are crucial for the reliability of the test setup.

3.2. Pre-formulated queries and cached result pages

Even for exactly the same task, the queries the searchers formulate vary vastly. Without control, task times with the interfaces may be different because the queries happened to be poor in one interface and good in the other. When the users' query formulation skills are not in the focus of the study, we have utilized *pre-formulated queries* and *cached result pages*. This way, the researcher can eliminate some variation from the test setup and be more confidential that the measured differences are due to the differences in the user interfaces.

The challenging aspect with the pre-formulated queries is to choose the query terms wisely – so that the results also apply outside of the experiment and that the queries resemble queries that the users would normally make. Having conducted numerous studies where the users formulate their own queries (Aula, 2003, 2005; Aula & Käki, 2003; Aula & Nordhausen, 2006), we have gained a good understanding on the terms the users typically select as query terms given the task description. Additional guidance for formulating realistic queries can be found from pages reporting common queries (e.g., http://www.google.com/press/zeitgeist.html), as well as from studies on transaction logs (Jansen & Pooch, 2001; Silverstein et al., 1999; Spink et al., 2001).

The main shortcoming with pre-formulated queries is that it is unfortunately possible for the researcher to deliberately choose the query terms so that her own interface wins the comparison. In fact, this may happen even for researchers with good intentions if they do not pay attention to the selection of query terms. Unfortunately, we have not come across with techniques to ensure fairness of the queries. One way to lessen the problem is to have the researchers always report the tasks they used along with the pre-formulated queries. Surprisingly, this essential information is often excluded from publications. Another shortcoming of pre-formulated queries is that they are not suitable for studies where the quality of the result set is in focus or when the interface specifically tries to aid users in query formulation.

To avoid the variance caused by the changing contents of databases and varying network delays, it is beneficial to locally save the result pages for the pre-formulated queries. With this approach we can be sure that all participants see exactly same set of results for a given query in a well controlled time-frame.

3.3. Time and data restrictions

From log studies, we know that the web searchers typically only evaluate one result page of ten results per query (Jansen & Pooch, 2001). However, in several of our early attempts to compare search interfaces, we noticed that users are often overly thorough in the test situation: they may spend several minutes evaluating the result list, which is not likely to happen in the real usage. Thus, we imposed a *time limitation* whereby the time the users had per task was limited to, for example, 1 min. According to our experiences, this limitation made the users' behavior closer to the real search behavior as seen from the transaction logs and thus, made our data more realistic.

The time limitation is not without problems: it may stress the participants as they may feel that they would have performed better given more time. Careful instructions emphasizing that whatever can be found in the given time is acceptable, can lessen the anxiety. In addition, the time restriction complicates the use of task time as a measure. Especially a tight limitation causes a ceiling effect and most of the task times will be exactly the given limit. However, this can be overcame by certain measures, such as search speed (discussed in Section 4.3).

Without proper control, the task time can easily contain activities that are not dependent on the quality of the search interface, such as, browsing through hyperlinks and reading the result pages. Additionally, going through information on web pages inevitably changes the users' understanding on the search task. If different users face different information during the test, their behavior may change unexpectedly and due to reasons that are not related to the search user interface. To eliminate these sources of variability, we have *restricted the users' access to web content* by disabling the links from the result page. This way, all the users see the same information (namely, result summaries) and the only factor that varies between the users is the interface. In these studies, the users have selected relevant-looking results by selecting a checkbox by the summary or by clicking on its title. Thus, the user interface looks realistic, but the behavior is modified to match the needs of the experimental setting.

4. Making comparisons easy with measures

An essential part of comparing search user interfaces is to utilize effective measures that focus on meaningful differences. In the context of comparing search user interfaces from the user-centered perspective, meaningful differences are, for example, differences in the user's performance (e.g., effectiveness and efficiency). In addition to that, measures dealing with the overall user experience are important to take into account, as the willingness to use the system is largely dependent on the user's perception of the system. Although we do feel that both performance and user experience factors need to be considered when evaluating search user interfaces, this paper purposefully concentrates on measures that capture the differences in the user's *performance* leaving the measures related to user experience out of discussion.

Even the performance measuring practices in the field vary considerably, which makes it difficult to understand the results from different studies and practically impossible to compare them. Good measures make comparisons within one study easy and also facilitate comparisons between different studies. For this purpose, we have found suitable measures from the earlier research and proposed a few new ones.

4.1. Interactive precision and recall

The traditional measures of *precision* and *recall* are the cornerstones of the research on information search. However, in user-centered studies, these measures are not effective as they do not consider the users' actions. In response to this problem, Veerasamy and colleagues (Veerasamy & Belkin, 1996; Veerasamy & Heikes, 1997) have proposed modified versions of the measures: *interactive recall* and *interactive precision*. Interactive recall measures the percentage of the relevant documents in the result set that were selected by the user whereas interactive precision states the proportion of relevant documents within the selected documents. In our studies, we have found these measures to be useful and to provide interesting information on the compared systems.

4.2. Immediate accuracy

When using web search engines, users typically open only a couple of documents per each query for closer inspection (Spink et al., 2001). *Immediate accuracy* captures the relevancy of these selections (Käki, 2004).

Immediate accuracy is a cumulative measure that states the proportion of cases where the users have found at least one relevant result by the *n*th result selection. Immediate accuracy of 85% by second selection means that in 85% of the cases (tasks) the users have found at least one relevant result by the time they have selected two result documents. Higher immediate accuracy means better success with this search style. It is noteworthy that immediate accuracy rarely reaches 100% meaning that not everyone will find a relevant result for each task.

We have employed this measure in several experiments (Käki & Aula, 2005; Käki, 2005a, 2006) and have found meaningful differences between interfaces with it. The results of this measure are easy to understand and seem to reflect the web search behavior well. In addition, being a proportional measure, the measure makes it easy to compare different systems.

4.3. Search speed

Users' search speed is one of the core measures in the user-centered evaluations and thus, we were surprised to see that there was no established measure for it. Previous studies frequently report raw measurements (e.g., task time and number of selected results) from the test setups, which are difficult to compare even within one study due to the lack of normalization. For example, it is not easy to see if System1 that takes on average, 10.10 min to find 16.44 answers is better or worse than System2, with which users can find 12.26 answers in 30.64 min (data from Pirolli, Schank, Hearst, & Diehl, 1996).

To overcome this problem, we proposed a proportional *search speed* that is measured in answers per minute (APM) (Käki, 2004). With search speed, the relationship of the previous two systems is evident, as System1 gets 1.62 APM and System2 gets 0.40 APM. In addition, these figures make it easy for the reader to understand the magnitude of speed that is achieved considering the constraints posed by the test setup. Although a normalized measure is used, the comparison between different studies is not trivial. Test setup has a vast effect on the results and the measure does not take this effect into account.

To enhance the comparisons even further, we introduced the measure of *qualified search speed* that considers the quality of the results in addition to time. It is measured in answers per minute for a relevancy category. Our example System1 could yield 1.32 irrelevant and 0.3 relevant results per minute. If System2 yielded 0.4 relevant results per minute, the qualified search speed reveals that System2 is better although the raw search speed suggested otherwise. In our studies, this measure has revealed that the source of performance difference is typically the increase in *relevant speed* while the *irrelevant speed* has remained nearly constant.

5. Applying the methods and metrics

To better understand how the proposed methods could be used in experimental comparisons of search user interfaces, we will present a few examples. The examples are based on our own studies where the methods were developed and applied.

5.1. Investigation of the readability of search results

Aula (2004) compared three different result presentation styles in effectiveness and efficiency (task time and error rate). All of the presentation styles contained exactly the same textual information, only the text layout and the use of bolding of query term occurrences varied between the conditions. The study used a *within-subjects design* where all the participants completed tasks with all three presentations styles. To make it possible to compare the task times between presentation styles, we prepared *three balanced task sets* (1, 2, and 3), where the corresponding tasks closely resembled each other. For example, for the task "Who is the principal of the University of Oulu" only the name of the university would change between task sets but the task would otherwise be the same. All of the tasks were simple fact-finding tasks with only one correct answer.

To control the variance caused by users' formulating different queries, *pre-formulated queries* were used. Again, the queries were formulated so that they closely resembled one another (e.g., "principal university oulu", "principal university helsinki", "principal university tampere"). These queries were submitted to a search engine and the first ten results were saved for each query. During the experiment, the participants interacted with the *cached result pages* instead of real web content to avoid variance caused by network delays or the index of the search engine changing between participants. Furthermore, we *restricted the users' access to web content*; the next task was presented immediately after the users clicked on the title of the result they thought contained the answer to the fact-finding task.

The combinations of task set (1, 2, 3) and presentation styles (A, B, and C) were counterbalanced between participants using *Latin Square*. In practice, this meant that participant 1 would see first style A with tasks from task set 1, then style B with task set 2, and finally, style C with task set 3. For participant 2, the order would be A2, B3, and C1 and for participant 3 A3, B1, and C2. For the next three participants, the order of *presentation styles* was B, C, and A (the order of task sets remained constant); and for participants 7–9, the presentation styles were presented in the order C, A, and B. The order of presenting the individual tasks was randomized.

In this study, the strict control over the setup made it possible to reliably show the differences in efficiency between the presentation styles.

5.2. Evaluation of a search result user interface solution

A series of studies were conducted to evaluate a novel user interface for accessing search results. This example applies to two of them (Käki & Aula, 2005; Käki, 2006) using the same experimental setup. The evaluated new interface (category interface) automatically categorizes search results. The aim of the categorization is to reveal major topics among the results and make it easier for the users to understand the result set and access topics that are relevant for them. The user interface lists 15 categories next to the result list. When the user selects a category, the result list is filtered to show only the results belonging to the selected category. The interface was hypothesized to let the users access moderately sized result sets (about 100–300 results) effectively and thus, overcome the inherent problems in result ranking when the user enters ambiguous (often short) queries.

In the evaluation of the proposed solution, we compared it with the *de facto* standard solution of ranked list of results (Google style result listing). Each participant used both user interfaces (*within-subjects design*), which required us to prepare *two balanced task sets*. The use of task sets was counterbalanced between the user interfaces and participants using *Latin Square*.

Each task was associated with a *pre-formulated query*. In these studies, this was seen to be mandatory to reliably compare the effect of the user interface rather than, for example, difference in users' query formulation skills. The queries were executed before the experiments and *query results were cached* to ensure the same stimuli for each participant.

During the experiment, we *denied the access to web content*, because the users' ability to evaluate and browse the actual web documents was not part of the phenomenon we studied. *Time for each task was limited* to 1 min. Although the restriction is fairly radical, it was seen both necessary and successful in this study. One minute was appropriate for this setup as the average number of collected results was similar as in an unrestricted situation observed by Aula and Nordhausen (2006).

In measuring the success of the compared user interfaces, we applied all the measures presented earlier. Reporting the *search speed* gave an idea of how fast the users were in the test situation. *Qualified search speed* revealed that the performance difference is a result of the users finding relevant results faster with the proposed category interface than with the conventional interface. *Interactive recall and precision* confirmed those findings by showing higher recall and precision measures for the category interface. In practice, this means that the users found a larger proportion of relevant results from the result set with the new interface and that the selected results contained less irrelevant results. Higher score in *immediate accuracy* indicated that the new user interface is also more effective in typical Web behavior where the first one or two good enough answers may be enough for the users.

6. Discussion

Researchers who have involved users in studies of search behavior or search systems can surely appreciate the complexity of the situation. When a human being is taken along into an evaluation, a number of uncontrolled variables come in to play. For example, we have seen enormous differences in the users' query formulation skills, their style of refining queries, and their habits and thoroughness of result evaluation. If such variables are not controlled in an experiment, it is impossible to draw reliable conclusions concerning the differences between the search interfaces being compared.

Table 1
Summary of the ways of controlling the complexity and enhancing comparisons via measures in user-centered search interface studies

Method	Used to	Pros	Cons
Ways of controlling the complexity of user-centered evaluations			
Time limitation	Make the users' behavior more realistic in the research situation that may make the searcher unnaturally thorough	Allows mimicking real search speed in a controlled fashion	Stresses participants, poses challenges for ecological validity
Pre-formulated queries & Cached results	Remove the variation due to the different search skills of users	Eliminates variation caused by differences in search skills	Poses challenges for ecological validity
Restricting access to content	Remove the time used for tasks that do not depend on the user interface quality	Makes the experiments shorter and more focused	Poses challenges for ecological validity
Measures for search success			
Interactive precision and recall	Describe the quality of the results the users find with the system	In line with classical measures of precision and recall	Ignores the time (efficiency) aspect
Search speed	Describe how fast the users can find results	Facilitates comparisons between search systems	Ignores the quality of the selected results
Qualified search speed	Describe how fast the users can find relevant or irrelevant results	Facilitates comparisons between search systems, accounts for the result quality	May not be applicable in every test setup
Immediate accuracy	Describe the success in the web style searching	Facilitates comparisons between search systems	Ignores the thoroughness of the search

In this paper, we shared our experiences in dealing with this complexity. We discussed various methods of controlling the experimental studies of search interfaces. Additionally, we presented measures for enabling meaningful comparisons of search user interfaces as summarized in Table 1.

We have successfully utilized the presented methods and metrics when comparing search user interfaces in the context of the Web. Furthermore, our experiences with the methods and metrics are based on studies where the interest has been mainly to support the users' evaluation of search results, rather than the search process on the whole. Due to these restrictions, the applicability of the methods to other contexts and situations is not trivial and needs to be considered case-by-case. However, in the following, we will aim at giving some general guidelines for applying the methods and metrics.

First, we expect most of the proposed methods and metrics to be applicable to other search contexts than Web, as well. We believe that within-subjects designs, balanced task sets, pre-formulated queries, cached result lists, limitation of the users' access to result documents, search speed, and qualified search speed can also be applied also to other search environments, such as digital libraries.

Second, the applicability of other methods or metrics, such as time limitations or immediate accuracy, should not be taken for granted. Limiting the users' time in the result evaluation phase seems to be an appropriate method in the context of the Web as Web searchers typically evaluate the results quickly. However, in the other contexts, the evaluation of search results may not follow this pattern. Immediate accuracy is a metric that specifically captures the behavior of web searchers and again, the applicability to other contexts may not be straightforward.

One must however, that all the presented methods have their shortcomings. As any method, they could even completely jeopardize the validity of the research results if applied inappropriately. Unfortunately there are no formal procedures for ensuring the valid usage of the methods but rather their suitability must be carefully considered for the research question at hand.

Our experiences with the methods and metrics are restricted in that we have focused on controlling the variability in the phase of results evaluation and those preceding it. Thus, in our studies, the users have not been allowed to formulate their own queries, spend as much time evaluating the results as they wish, or read the documents the results point to. Consequently, the studies have not given any new information about these phases of search. Future research needs to address the questions of which controlling methods are suitable when the aim is, for example, to experimentally study interface solutions aimed at facilitating initial query formulation.

References

Anick, P. (2003). Using terminological feedback for web search refinement: A log-based study. In *ACM SIGIR conference on research and development in information retrieval (SIGIR'03)* (pp. 88–95).

Aula, A. (2003). Query formulation in web information search. In *Proceedings of IADIS international conference WWW/Internet 2003* (pp. 403–410).

Aula, A. (2004). Enhancing the readability of search result summaries. In A. Dearden & L. Watts (Eds.) *Proceedings of the conference HCI 2004: Design for life* (pp. 1–4).

Aula, A. (2005). Older adults' use of web and search engines. *Universal Access in the Information Society, 4*(1–2), 67–81.

Aula, A., Jhaveri, N., & Käki, M. (2005). Information search and re-access strategies of experienced web users. In *Proceedings of WWW 2005* (pp. 583–592).

Aula, A., & Käki, M. (2003). Understanding expert search strategies for designing user-friendly search interfaces. In *Proceedings of IADIS international conference WWW/Internet 2003*, (Vol. II, pp. 759–762).

Aula, A., & Käki, M. (2005). Less is more in web search interfaces for older adults. *First Monday, 10*(7).

Aula, A., & Nordhausen, K. (2006). Modeling successful performance in web search. *Journal of the American Society for Information Science and Technology, 57*(12), 1678–1693.

Brajnik, G., Mizzaro, S., & Tasso, C. (1996). Evaluating user interfaces to information retrieval systems: A case study on user support. In *Proceedings of the 19th International conference on research and development in information retrieval (SIGIR'96)* (pp. 128–136).

Dumais, S., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'01)* (pp. 277–284).

Ford, N., Miller, D., & Moss, N. (2005). Web search strategies and human individual differences: Cognitive and demographic factors, Internet attitudes, and approaches. *Journal of the American Society for Information Science and Technology, 56*(5), 741–756.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'00)* (pp. 345–352).

Hertzum, M., & Frøkjær, E. (1996). Browsing and querying in online documentation: A study of user interfaces and the interaction process. *ACM Transactions on Computer–Human Interaction, 3*(2), 136–161.

Heimonen, T., & Jhaveri, N. (2005). Visualizing query occurrence in search result lists. In *Proceedings of the 9th international conference on information visualization (IV'05)* (pp. 977–882).

ISO 9241-11. *Ergonomic requirements for office work with visual display terminals – part 11: Guidance on usability (ISO/IEC 9421-11: 1998)*.

Jansen, B. J., & Pooch, U. (2001). A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology, 52*(3), 235–246.

Jones, W., Bruce, H., & Dumais, S. (2001). Keeping found things found on the Web. In *Proceedings of the Tenth international conference on information and knowledge management 2001* (pp. 119–126).

Käki, M. (2004). Proportional search interface usability measures. In *Proceedings of NordiCHI 2004* (pp. 365–372).

Käki, M. (2005a). Optimizing the number of search result categories. In *Proceedings of the SIGCHI conference on human factors in computing systems, CHI 2005* (pp. 1517–1520).

Käki, M. (2005b). Findex: Search result categories help users when document ranking fails. In *Proceedings of the SIGCHI conference on human factors in computing systems, CHI 2005* (pp. 131–140). ACM Press.

Käki, M. (2006). fKWIC: Frequency based keyword-in-context index for filtering web search results. *Journal of the American Society for Information Science and Technology, 57*(12), 1606–1615.

Käki, M., & Aula, A. (2005). Findex: Improving search result use through automatic filtering categories. *Interacting with Computers, 17*(2), 187–206.

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 of Information Science, 51*(6), 576–581.

Paek, T., Dumais, S., & Logan, R. (2004). WaveLens: A new view onto Internet search results. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'04)* (pp. 727–734). New York: ACM Press.

Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the ACM CHI'96* (pp. 213–220). ACM Press.

Savage-Knepshield, P. A., & Belkin, N. J. (1999). Interaction in information retrieval: Trends over time. *Journal of the American Society for Information Science and Technology, 50*(12), 1067–1082.

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum, 33*(1), 6–12.

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.

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology, 52*(3), 226–234.

Su, L. T. (1994). The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science and Technology*(April), 207–217.

Tague, J., & Schultz, R. (1988). Some measures and procedures for evaluation of the user interface in an information retrieval system. In *Proceedings of the 11th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'88)* (pp. 371–385).

Topi, H., & Lucas, W. (2005). Mix and match: Combining terms and operators for successful Web searches. *Information Processing and Management, 41*, 801–817.

TREC interactive track guidelines. (1999). <http://www-nlpir.nist.gov/projects/t8i/spec.html> (checked Sep. 5th, 2006).

TREC interactive track guidelines. (2000). <http://www-nlpir.nist.gov/projects/t9i/spec.html> (checked Sep. 5th, 2006).

TREC interactive track guidelines. (2002). <http://trec.nist.gov/data/t11_interactive/guidelines.html> (checked Sep. 5th, 2006).

Veerasamy, A., & Belkin, N. (1996). Evaluation of a tool for visualization of information retrieval results. In *Proceedings of the annual international ACM/SIGIR'96 conference* (pp. 85–92). ACM Press.

Veerasamy, A., & Heikes, R. (1997). Effectiveness of a graphical display of retrieval results. In *Proceedings of the Annual International ACM/SIGIR'97 Conference* (pp. 6–15). ACM Press.

White, M. D., & Iivonen, M. (2001). Questions as a factor in Web search strategy. *Information Processing and Management, 37*(5), 721–740.

White, R. W., Ruthven, I., & Jose, J. M. (2002). Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *Proceedings of the 27th annual international conference on research and development in information retrieval (SIGIR'02)* (pp. 57–64). New York: ACM Press.

White, R. W., Ruthven, I., & Jose, J. M. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing and Management, 39*(5), 707–733.

Woodruff, A., Faulring, A., Rosenholtz, R., Morrison, J., & Pirolli, P. (2001). Using thumbnails to search the Web. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'01)* (pp. 198–205). New York: ACM Press.