3rd International Workshop on Perceptual Quality of Systems (PQS 2010)
6-8 September 2010, Bautzen, Germany

User Experiments with Search Services: Methodological Challenges for Measuring the Perceived Quality

Katrin Lamm[1], Thomas Mandl[1], Christa Womser-Hacker[1], Werner Greve[2]

[1] Information Science, University of Hildesheim, Germany
[2] Psychology, University of Hildesheim, Germany
{lammka, mandl, womser, wgreve} @uni-hildesheim.de

Abstract

For many people, search engines are a crucial entry point for their online activities today. People use search services to find relevant information of their interest on the web. Measuring the quality of search systems is a challenging task. User expectations have been proposed as one possible parameter for user satisfaction. The so-called confirmation/disconfirmation (C/D) paradigm, a widely used model to describe customer satisfaction, was used to predict the reaction of search engine users. Two studies were conducted to examine how prior expectations of users affect their perceptions of system quality of search engines. This paper describes the initial test designs, compares the methodical approaches and presents a concise summary of statistical results obtained using ANOVA.

Index Terms: interactive information retrieval, user expectations, user satisfaction, confirmation/disconfirmation paradigm

1. Introduction

Searching for information on the internet is an iterative process. When searching for information users normally do not invest their time into working out exactly what they are trying to find. More likely, they will apply multiple queries one after the other to retrieve useful information by chance rather than issuing expertise queries. Therefore, real-life searching often consists of a sequence of preferential short queries. Nevertheless, Järvelin shows that individually ineffective queries may be surprisingly effective if evaluated in the perspective of the overall session performance [1]. Fast internet connections and increased computing power have eliminated the need of getting a query right on the first try. These observations seem to suggest that the so-called Cranfield paradigm [2], which refers to a batch- and system oriented way of IR evaluation, may need to be complemented by a user-oriented approach of evaluating the effectiveness of IR systems.

In recent years, researchers start to think outside the traditional paradigm [3]. By explicitly taking into account the users, their tasks, and the context in which they work, the researchers aim to gain a deeper understanding of how users respond to the perceived quality of search engine results. There has been a growing concern on whether the results of batch and user experiments return the same results [4, 5, 6, 7, 8, 9]. The central question in this context is whether better systems enable users to find more relevant documents and if user satisfaction does correlate with result list quality? So far, comparatively few studies focus on the satisfaction aspect of this question [10, 11, 12]. Both the experimental setups and results of these studies offer a rather heterogeneous picture and thus hamper comparison among studies.

In this paper, we stress the need to include user expectations into the user-centered evaluation and we discuss the challenges involved in such experiments. The users' overall evaluation of search engine results not only depends on the binary decision of relevant or not relevant, it is also determined by prior expectations. If we take the query *dollar exchange rate* as an example, the exact same query may return highly satisfying results to user A but totally inappropriate results to user B. This discrepancy may be due to different expectations. Maybe user A was looking for the *dollar exchange rate today*, while user B was interested in the *dollar exchange rate history*. In the first case probably one reliable reference will be sufficient (question answering style information need). In the other case, however, the user might wish to find quite a lot references to be able to compare amongst them (ad-hoc style information need). One could argue that the query is not well defined, but that is how users interact with search engines and therefore it has to be dealt with in a holistic way. At this point it is important to note that the individual information need is only one factor which influences user expectations. Further factors may include the prior knowledge and experience of the user, the corporate communication of the search engine provider as well as the word of mouth between friends and acquaintances.

In traditional customer satisfaction research it has long been known that there is a close link between a customer's expectation and perception of quality. A widely used model to describe customer satisfaction is the so-called confirmation/ disconfirmation (C/D) paradigm, which explains the creation of (dis)satisfaction by the confirmation or disconfirmation to expectations.

In this paper we report two studies investigating users' perceptions of the quality of search services and the effect of prior expectations on discerning search results. This paper describes the initial test designs, compares the methodical approaches and presents a concise summary of statistical results obtained using ANOVA.

2. Fundamentals of IR Evaluation

Information retrieval (IR) deals with the search for information and the representation, storage and organisation of knowledge. Information retrieval is concerned with search processes in which a user needs to identify a subset of information which is relevant for his information need within a large amount of knowledge.

The user is in the center of the information retrieval process. Nevertheless, most research tends either to be more user-oriented or more algorithm and system-oriented. User-oriented research tries to pursue a holistic view of the process whereas system-oriented research is concerned with measuring the effect of system components. The information retrieval process is inherently vague. In most systems, documents and queries traditionally contain natural language. The content of these documents needs to be analyzed, which is a hard task for computers. Robust semantic analysis for large text collections or even multimedia objects has yet to be developed. Therefore, text documents are represented by natural language terms mostly without syntactic or semantic

10.21437/PQS.2010-10

context. These keywords or terms can only imperfectly represent an object because their context and relations to other terms are lost.

As information retrieval needs to deal with vague knowledge, exact processing methods are not appropriate. As a consequence, the performance of a retrieval system cannot be predicted but must be determined in evaluations. Evaluation plays a key role in information retrieval. Evaluation needs to investigate how well a system supports the user in solving his knowledge problem.

Obviously, a good system should satisfy the needs of a user. However, the users' satisfaction requires good performance in several dimensions. The quality of the results with respect to the information need, system speed and the user interface are major dimensions. Due to the difficulties of assuming a holistic perspective in the evaluation, most often a system-oriented approach was followed in IR research. The adopted evaluation scheme tries to ignore subjective differences between users in order to be able to compare systems and algorithms. Relevance judgments are given out by experts who evaluate the relevance of a document independent of subjective influences. This approach is called the *Cranfield-Paradigm* after one early evaluation study [13].

The most important basic measures are recall und precision. Recall indicates the ability of a system to find relevant documents, whereas precision measures show how good a system is in finding only relevant documents without ballast. Recall is calculated as the fraction of relevant documents found among all relevant documents, whereas precision is the fraction of relevant documents in the result set. The recall requires knowledge of all the relevant documents in a collection that could never be put together in any real world collection. The number of known relevant documents is usually used to calculate the value. Both measures are set oriented. However, most current systems present ranked results. In this case, a recall and precision value pair can be obtained for each position on the ranked list taking into account all documents from the top of the list down to that position. Plotting these values leads to the recall-precision graph. The average of precision values at certain levels of recall is calculated as the mean average precision (MAP), which expresses the quality of a system in one number.

Evaluation initiatives compare the quality of systems by determining the mean average precision for standardized collections and topics like descriptions of information needs. The relevant documents for the topics are assessed by humans who work through all the documents in a pool. The pool is constructed from the results of several systems and ultimately limits the number of relevant documents which can be encountered.

Important experiments are carried out within evaluation initiatives. The three major evaluation initiatives are historically connected. The Text Retrieval Conference (TREC, trec.nist.gov) was the first large effort which started in 1992, followed by the NTCIR (NII-NACSIS Test Collection for IR Systems, research.nii.ac.jp/ntcir/) Project where the first workshop was held in 1999 and the Cross Language Evaluation Forum (CLEF, clef-campaign.org/) which started in 2000.

User based evaluations confront test users with the results of search systems and let them solve information tasks given in the experiment. In such a test, the performance of the user can be measured by observing the number of relevant documents she or he finds. This measure can be compared to a gold standard of relevance for the search topic to see if the perceived performance correlates with an objective notion of relevance defined by a juror. In addition, the user can be asked about his satisfaction with the search system and its results. Only recently, user based experiments have gained more attention. In 2006, Al-Maskari et al. [14] conducted a study in association with a submission to iCLEF2006. iCLEF is the interactive track of CLEF. In 2006, image retrieval was selected as the central theme of this track. Pictures from the photo sharing community FLICKR were used as data collection. Several statistical effectiveness measures were used to evaluate the system performance. In addition, user performance was measured via task-specific modifications of recall and precision. The authors also asked participants for their satisfaction with the usefulness, accuracy and coverage of the search results. Participants rated their satisfaction on a 3-point scale (1 = very satisfied; 0.5 = partially satisfied; 0 = not satisfied). No direct relationship between system performance and user performance resp. satisfaction could be confirmed within the scope of this user study [14].

In another study, Al-Maskari et al. [10] conducted a similar experiment on the satisfaction of users. This time, participants directly searched in the internet search engine Google . The purpose of this study was to determine whether a correlation between the effectiveness of Google results quantified by statistical measures and user satisfaction could be established. The latter was again assessed using a 3-point scale. Instead of the satisfaction with the usefulness this time the users' satisfaction with the ranking of results was requested. In this experiment a correlation between system performance and user satisfaction could be proven [10].

Huffman and Hochster [11] pursued a slightly different approach. The experiment was also based upon the Google search engine. The test users were given real queries already submitted to Google as information needs. Within this user satisfaction study, one group of test users rated the results in terms of their relevancy and another group in terms of their satisfaction. Subsequently, the latter group was asked to submit the queries again and then act as if they really had this information need. In order to investigate the relationship between system performance and user satisfaction Huffman and Hochster contrasted the relevance judgments for the top three search results of the first query of each session with the user's final satisfaction.

In the years 2000 and 2001 Turpin and Hersh [9] carried out two user studies on the question whether batch and user evaluations lead to comparable results. The first study was performed within the framework of the TREC-8 interactive track and consisted in an instance recall task. The second study was carried out within the TREC-9 interactive track and consisted of a question-answering task. For both experiments, the authors used two search systems with different MAP performance (TREC-8: 0.27 vs. 0.32 MAP; TREC-9: 0.27 vs. 0.35 MAP). An influence of the system performance on the user performance could neither be observed for the instance recall nor for the question-answering task. Merely for one measure, a relation could be observed [9].

The two following studies investigated the influence of different levels of system quality on the user performance by using artificially constructed result lists. That way, the system performance can be better controlled. Whereas Allan et al. [4] adopted bpref to measure the system performance, Turpin and Scholer [8] adopted MAP. Turpin and Scholer found it especially important to use simple search tasks. Their precision-oriented task consisted in finding a document that is relevant to an information need (time was measured). The recall-oriented task consisted in finding as many relevant documents as possible within a given time limit of five minutes. Their results show a weak effect for the recall-oriented task [8]. Allan et al. employed a passage retrieval task in which subjects were required to find, highlight and label all facets of the answer to a given information need. Their results

show noticeable differences in the user performance at certain levels of bpref [4].

In another study, Scholer and Turpin [6] analyzed the concept of an individual's relevance threshold in relation to the system performance. The aim of this study was to further understand the mismatch in the results between batch and user evaluations. The diversity of the user's relevance criteria constituted the starting point for establishing a study design that opens up the possibility to investigate their relevance judgments more precisely. Therefore, similar to their previous study from 2006, Scholer and Turpin artificially constructed result lists to simulate several individual search systems with three different relevance criteria: One pair of inferior and superior systems with a strict irrelevance criterion, one pair with a strict relevance criterion and one pair that represents a mix of the two. The user performance measure was the time users took to find relevant documents with the particular system. Findings suggest that different users adopt different relevance criteria. As a consequence, a large variance in system performance which can be typically observed in batch experiments does not always occur in user experiments [6].

Several psychological factors regarding the user experience when searching for information have not been considered yet. In addition, there are many methodological issues involved in user experiments.

3. Theoretical and Empirical Issues Concerning Satisfaction

User oriented studies for information retrieval systems have neglected research on customer satisfaction from other areas. A great deal of research on satisfaction has been done in the field of marketing research. In marketing literature the confirmation or disconfirmation of prior expectations is considered to be an important factor contributing to customer satisfaction [15, 16]. This observation has led to the illustrative C/D paradigm, according to which, a customer is satisfied if target and actual performance match. This state of mind is also referred to as confirmation. In the case of disconfirmation of expectations there are two possible outcomes. Customer expectations can be surpassed also referred to as positive disconfirmation, or disappointed, also referred to as negative disconfirmation.

With respect to the relationship between expectations and satisfaction we want to especially draw your attention to a study conducted by Szajna and Scamell in 1993 [17], as this study both in terms of content and methods used is very close to our experiments presented in this paper. Szajna and Scamell investigate the effects of user expectations in the context of an information system. Unlike in our experiments, cognitive dissonance theory was used to predict the reaction of users. This theory, developed by Festinger [18], assumes that individuals tend to feel the need for cognitive harmony. To counteract cognitive dissonance individuals tend to either lower their expectations or raise their perceptions. In a longitudinal experiment, Szajna and Scamell control the expectations of 159 business students to be high, moderate or low in order to investigate the relation between expectations, user satisfaction and performance. Results show a connection between the realism of user expectations and perceptions of the information system, but not the user performance [18]. Furthermore, this study revealed that user expectations tend to *wear of* over time.

4. Methods

In order to investigate the association between user expectation and system performance on the one hand and user satisfaction and performance on the other hand, two laboratory wizard-of-oz experiments were conducted. In everyday IR practice there are certainly many factors important for the success and the course of action of a search process. As a consequence, a multifactorial design was employed in both studies. Both of our designs involve the manipulation of the system quality as well as the user expectation as independent variables in order to determine their influence on a set of dependent variables. The dependent variables in both cases include a user satisfaction questionnaire as well as some effectiveness measures which intend to quantify user performance.

4.1. Independent Variables

The independent variables were the system quality and the realism of user expectations. To control the system performance, as in previous studies [4, 6, 8, 12], artificially constructed result lists have been used. For each search topic, two different result lists were created, one to simulate a low and one to simulate a high quality search engine. The participants were unaware that the test system actually only simulated the real-life search with a web search engine. User expectations were manipulated to be comparatively low or comparatively high through the test instructions prior to the tests. In order to manipulate the expectations in the first study, the system was in one case introduced as an expensive professional search system and in the other case as a student project with unknown quality. For a more detailed description of this study, see [19]. In order to improve the expectation manipulation, in the second study, a different manipulation strategy was used where each participant was asked to use two supposedly different search engines. To manipulate their expectations the users were told that recent system performance tests had shown that system A is better than system B. As in the study by Kelly et al., the suspected search engines were named after colors to help the subjects differentiating between them [12]. Independently of this manipulation the system performance has been varied as explained above.

Table 1. *Experimental design study 1.*

		System	
		good	Bad
Expectation	low	Group 1	Group 2
	high	Group 3	Group 4

This resulted in a study design with four different treatment groups, differing in their expectation and the actual system quality for the first study and eight different treatment groups for the second study. The fact that each participant used both suspected systems led to the four additional groups in the latter case. Table 1 and 2 show the experimental designs of both studies. The distribution of participants was done randomly to ensure that the results obtained are due to the experimental manipulation and participants were unaware of the different experimental conditions.

Table 2. *Experimental design study 2.*

		System			
		bad-bad	bad-good	good-good	good-bad
Expectation	low-high	Group 1	Group 2	Group 3	Group 4
	high-low	Group 5	Group 6	Group 7	Group 8

In both studies each subject participated in only one treatment. The tables read as follows: Subjects of group 1 in study 1 got the low expectation manipulation instruction, but were presented with the comparatively bad search engine result pages. In contrast expectation manipulation and actual system performance matched for group two. Subjects of group 1 in study 2 got the instruction they were to be using the bad system (low expectation) before the good system but actually both times were presented with the bad search engine result pages.

In line with the C/D paradigm, it was assumed that (1) subjects with unrealistically low expectations would perceive positive disconfirmation of their expectations and potentially be satisfied, (2) subjects with realistically low or realistically high expectations would have their expectations confirmed and potentially be satisfied and (3) subjects with unrealistically high expectations would perceive negative disconfirmation of their expectations and potentially be dissatisfied. The realism of expectations in this particular case is defined as the match or mismatch of expectation and system manipulation.

4.2. Dependent Variables

The dependent variables in our experiments include user satisfaction and performance. As pointed out earlier interactive IR evaluation is a relatively new area of activity. Therefore, the methods used to measure satisfaction and performance are not yet standardized through common research practice.

Sections 3 and 4 have shown that user satisfaction is a complex construct that involves the interaction of multiple elements such as expectations, knowledge, experience and perceptions. In our experiments, satisfaction was measured with questionnaires developed specifically for the two studies. Both questionnaires comprised statements and questions that covered the following components of satisfaction: ease of use, efficiency, output display, precision, ranking of results, result quality and reuse probability. In contrast to the first study, the questionnaire used in the second study was improved and extended by the addition of the translated and adapted end-user computing satisfaction (EUCS) instrument by Doll and Torkzadeh [20].

The user performance was measured in terms of completeness (recall) and accuracy (precision) of documents saved as well as the time taken to find the first relevant document (TIME). Table 3 summarizes the performance measures used to determine the search success of our participants.

Table 3. *Performance measures study 1 and 2.*

Measure	Description	Equation
RD	Number of relevant documents found	RD
UR	Number of relevant documents found divided by total number of relevant documents in result list	$\frac{RD}{TRL}$
TIME	Time to find first relevant document	TIME
UP	Number of relevant documents found divided by total number of documents saved as relevant by user	$\frac{RD}{TRS}$
PCP	Number of relevant documents found divided by total number of documents selected as possibly relevant by user	$\frac{RD}{TPR}$

It is important to note, that the number of relevant documents found (RD) in this context refers to the number of documents correctly identified as relevant as compared to the relevance assessments used as ground truth. Whereas user recall (UR) is a measure of completeness, user precision (UP) can be seen as a measure of accuracy. Therefore, in the first case RD is related to the total number of relevant documents in the result list (TRL) and in the latter case RD is related to the total number of documents saved as relevant by the user (TRS). Pre-Click-Precision (PCP) is a measure of the user's first impression of the document result list. For that reason RD is related to the total number of documents selected as possibly relevant by the user (TPR).

4.3. Subjects

The sample of the first user test in 2008 consists of 89 female students, with each subject randomly assigned to one of the four treatment groups. Because of gender differences in searching [21], only female subjects were invited for this study. The sample of the second user test in 2009 includes 118 participants, 90 female and 28 male students of our University. Subjects of both studies had some searching experience and tended to use the internet on average 16.7 (SD ± 12.8) hours per week in the first study and 18.4 (SD ± 10.4) hours per week in the second study.

4.4. Tasks

One commonly stated goal of interactive IR research is to mirror a real usage situation as close as possible. Thus, it was important that the appearance of the test systems, simulated in the two experiments, can be compared to current internet search engines and the tasks reflect real life contexts. Furthermore, the handling of the systems had to be easy in order to not distract the users from the actual task.

In both tests, subjects were instructed to find relevant documents given a well-defined information need. Users could perform a limited number of actions in the user interfaces. They could browse through a list of result documents and evaluate them based on a representation by title and snippet. Users could then select documents by clicking. Subsequently, the web document appeared within the interface of the search engine, users could read them and decide if they were relevant or not. In order to find relevant documents they did not have to read through all documents in the result lists but regard only those that seemed relevant to them after their first impression based on title and snippet. Once a document was selected, they had to decide whether or not the document was relevant by selecting the associated checkboxes. Judgments and timestamps for all interactions with the test systems were

recorded in a log file. The time allotted to complete each task was ten minutes, although subjects could terminate the search early if they felt they had completed the task. This was done to ensure more natural behavior and hence help minimize the artificiality of the test situation.

4.5. Operational Differences

The major operational differences between the two studies were the following.

As the expectation manipulation of the first study did not appear to affect the independent variables in a statistically significant manner, one goal of the second study was to improve the manipulation of user expectations. The idea behind the new experimental design was that the comparison of two different systems can help to mediate the effect of the expectation manipulation. A manipulation check was accomplished by asking the participants to recall which search engine scored better in recent system performance tests.

Whereas the test instructions in the first study were given in writing and the test system was installed on a local computer, in the second study the instructions were given on video and the system was installed on a central server. This made it possible to test multiple participants at the same time.

Another difference lies in the test collections used for the experiments. Using a sub-set of a standard newswire test collection developed by CLEF in the first study had the advantage that a set of documents, topics and relevance assessments already existed. For reasons of document actuality and to have a high level of practical relevance it was decided to develop a new corpus of web documents for the second study.

The experimental procedure also differed slightly from experiment to experiment. Due to the experimental design, in the first user test subjects performed three searches on three different topics always being presented with result lists of the same system performance manipulation. In contrast, in the second user test subjects performed only two searches on two different topics but with two supposedly different systems and being presented with result lists of varying system performance depending on the experimental condition they were in. To control for possible order effects such as learning and fatigue in both experiments the task order was counterbalanced across subjects and treatment groups. For the second experiment this means that half of the group completed the first topic with the first system and the second topic with the second system and half of the group did it the other way around and thus completed the second topic with the first system and the first topic with the second system. The design also determines the time the satisfaction questionnaires were handed out to the participants. While in the first experiment it was handed out at the very end of the test, in the second experiment subjects were asked to complete one questionnaire at the end of each system use.

The major difference between the two studies concerns the search behavior established through the experimental design. In the first study the typical real-world iterative search behavior was disabled. In order to provide each subject with the same result lists, the queries were predetermined and participants were asked not to reformulate them. In the second study user behavior that goes beyond simply scanning a list of search results was supported by allowing participants to issue several queries relating to each topic. As in the study by Turpin and Scholer, for each query submitted in a session, a random result list from a collection of possible lists was returned and identical queries within a session induced the same lists again [8]. The link color of already seen documents was not changed during the second user test in order to confirm the impression of a real search system. In the case that a document was selected for the second time, the previous judgment would be selected in the checkbox as long as the participant did not change the mind.

5. Results

For the first time, the C/D paradigm has been applied to IR and therefore user expectation has been modeled and analyzed in a user experiment. In both studies we found evidence that the expectation of users is dynamic and context dependent. Some results of the first study suggest that prior expectations have an effect on the satisfaction as predicted by the C/D paradigm. However, there were no statistically significant differences. The results of the second study did not corroborate this observation.

With respect to performance expectations results of the second study also imply that user expectations do not seem to have a lasting effect. This can be seen from the fact that a significant correlation between user expectation and user satisfaction could only be found for the first search task. For the second search task system performance seems to prevail and significant correlations can only be found in the case of system performance. Our results therefore seem to suggest that expectations can be *overwritten* by performance experience. This conclusion is supported by the findings of Szajna and Scamell, who reported that unrealistic expectations tend to *wear of* over time [17]. This effect might also be the reason why we did not observe significant differences regarding the user expectations in the first experiment. Maybe the experimental procedure in which subjects completed the satisfaction questionnaire at the very end of the test has resulted in the fact that the manipulated expectations have become more realistic through the use of the test system over three search tasks. Thus perhaps results would have been different for the first experiment if a questionnaire would have been handed out after the completion of each task instead.

Apart from that the results show that self-reported relevance in user studies is highly context dependent. This is evidenced by the fact that users significantly seem to relax their relevance criteria as soon as they begin to use the lower quality search engine.

For further results of the first study, also see [19], more findings of the second study will be discussed elsewhere.

6. Discussion

Several challenges for user experiments in IR were better solved in the second study. On the one hand we can surely say that the realism of the experimental design, or the extent to which the design mimics the natural search experience of users, has been improved from the first to the second user test. On the other hand, these operational improvements resulted in a more complex and less predictable experimental setting, because they may introduce several possible confounding, uncontrolled variables into the analysis.

In this regard, one important practical question is how to measure the users' search performance if each user performs a different number of searches within a session and also receives different result lists. While we had to compare the searcher performance based on two result lists per topic for the first study, we had to deal with randomly generated result lists in the second study. In the following, we briefly explain how the user performance measures presented in section 4.2 can be applied to this new situation. As the search behavior of users could vary quite a lot through the new design, it was decided to measure user performance as overall session performance. That is to say that the document sets needed for the user-based

performance calculation include the documents from all queries issued during a single session, duplicates were counted once. Hence the total number of relevant documents in the result list (TRL) in the second experiment refers to the total number of unique relevant documents in the result lists presented within a single session. The same applies to RD, TRS and TPR, only unique documents were counted.

The next important question concerns the way in which the statistical analysis of the test data may be carried out. Here again the experimental design has major implications for the methodical procedure. Between-group comparisons were performed by ANOVA. Corresponding to the study designs, system performance and user expectation formed the independent variables. The more subjective user perception and the more objective user performance were included as dependent variable. Because the second study, in contrast to the first study, involved the use of two supposedly different systems, two such analyses had to be done, one for each task-system pair. It is important to note in this context that generally there are two ways to evaluate the second experiment statistically. The first possibility is to treat each task-system pair individually, thereby accepting the fact that the experience of using the first system may have an additional confounding effect on the results of the second system. Another option is to analyze the first task-system pair individually, but take the first experience into account while analyzing the second task-system pair. As we do believe that prior experiences can influence future search behavior and perceptions of search results, we chose the second option. Hence, three factors were used for ANOVA: user expectation, system performance of first and second task. The user expectation did not have to be included twice, because it was clear that the expectation manipulation was high for the second task if it was low for the first and vice versa.

7. Conclusions

In a nutshell, it can be said that interdisciplinary collaboration offers a promising approach for interactive IR evaluation. The results of both studies highlight the need to understand and operationalize the origins of user expectations in order to closely mimic the natural search process. However, despite this, we still do not fully understand how these factors work in concert. Therefore in future studies, we intend to further elaborate the concept of user expectations. Future research should work to establish reliable methods to measure user satisfaction and performance in IR contexts. One important area for future research would be the development of an instrument to measure user expectations. The construction of such an instrument would help to ensure that the expectation manipulation and the expectations thereby created do indeed alter the users' behavior and perceptions.

8. References

[1] Järvelin, K., "Explaining User Performance in Information Retrieval: Challenges to IR Evaluation", Proc. International Conference on the Theory of Information (ICTIR) '09, 289-296, Springer, 2009.

[2] Mandl, T., "Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance", Informatica, 32(1):27-38, 2008. Online: http://www.informatica.si, accessed on May 25, 2010.

[3] Kamps, J., Geva, S., Sakai, T., Trotman, A. and Vorhees, E., "Report on the SIGIR 2009 Workshop on the Future of IR Evaluation", SIGIR Forum, 43(2):13-23, 2009.

[4] Allan, J., Carterette, B. and Lewis, J., "When Will Information Retrieval Be "Good Enough"?", Proc. Annual ACM SIGIR Conference on Research & Development on Information Retrieval '05, 433-440, ACM, 2005.

[5] Al-Maskari, A., Sanderson, M. and Airio, E., "The Good and the Bad System: Does the Test Collection Predict Users' Effectiveness?", Proc. Annual ACM SIGIR Conference on Research & Development on Information Retrieval '08, 59-66, ACM, 2008.

[6] Scholer, F. and Turpin, A., "Relevance Thresholds in System Evaluations", Proc. Annual ACM SIGIR Conference on Research & Development on Information Retrieval '08, 693-694, ACM, 2008.

[7] Smith, C. L. and Kantor, P. B., "User Adaptation: Good Results from Poor Systems", Proc. Annual ACM SIGIR Conference on Research & Development on Information Retrieval '08, 147-154, ACM, 2008.

[8] Turpin, A. H. and Scholer, F., "User Performance versus Precision Measures for Simple Search Tasks", Proc. Annual ACM SIGIR Conference on Research & Development on Information Retrieval '06, 11-18, ACM, 2006.

[9] Turpin, A. H. and Hersh, W., "Why Batch and User Evaluations Do Not Give the Same Results", Proc. Annual ACM SIGIR Conference on Research & Development on Information Retrieval '01, 225-231, ACM, 2001.

[10] Al-Maskari, A., Sanderson, M. and Clough, P., "The Relationship between IR Effectiveness Measures and User Satisfaction", Proc. Annual ACM SIGIR Conference on Research & Development on Information Retrieval '07, 773-774, ACM, 2007.

[11] Huffman, S. B. and Hochster, M., "How Well does Result Relevance Predict Session Satisfaction?", Proc. Annual ACM SIGIR Conference on Research & Development on Information Retrieval '07, 567-574, ACM, 2007.

[12] Kelly, D., Fu, X. and Shah, C., "Effects of Rank and Precision of Search Results on Users' Evaluations of System Performance", TR-2007-02, University of North Carolina, 2007. Online: http://sils.unc.edu/research/publications/reports/TR-2007-02.pdf, accessed on May 31, 2010.

[13] Robertson, S., "On the History of Evaluation in IR", Journal of Information Science, 34(4): 439-456, 2008.

[14] Cadotte, E. R., Woodruff, R. B. and Jenkins, R. L., "Expectations and Norms in Models of Consumer Satisfaction", Journal of Marketing Research, 24(3): 305-314, 1987.

[15] Al-Maskari, A., Clough, P. and Sanderson, M, "Users' Effectiveness and Satisfaction for Image Retrieval", Proc. Workshop Information Retrieval 2006 of the Special Interest Group Information Retrieval (FGIR) '06, 84-88, 2006. Online: 2070116.pdf, accessed on June 8, 2010.

[16] Patterson, P. G., "Expectations and Product Performance as Determinants of Satisfaction for a High-Involvement Purchase", Psychology and Marketing, 10(5): 449–465, 1993.

[17] Szajna, B. and Scamell, R. W., "The Effects of Information System User Expectations on Their Performance and Perceptions", MIS Quarterly, 17(4): 493-525, 1993.

[18] Festinger, L., "A Theory of Cognitive Dissonance", Stanford: Stanford Univ. Press, 1957.

[19] Lamm, K., Mandl, T., Womser-Hacker, C. and Greve, W., "The Influence of Expectation and System Performance on User Satisfaction with Retrieval Systems", Proc. International Workshop on Evaluating Information Access (EVIA) '10 (to appear)

[20] Doll, W. J. and Torkzadeh, G., "The Measurement of End-User Computing Satisfaction", MIS Quarterly, 12(2): 259-274, 1988.

[21] Large, A., Beheshti, J. and Rahman, T., "Gender Differences in Collaborative Web Searching Behavior: An Elementary School Study", Information Processing and Management, 38(3): 427-443, 2002.