The Influence of Expectation and System Performance on User Satisfaction with Retrieval Systems

Katrin Lamm
University of Hildesheim
Marienburger Platz 22
31141 Hildesheim
katrin.lamm@uni-hildesheim.de

Thomas Mandl
University of Hildesheim
Marienburger Platz 22
31141 Hildesheim
mandl@uni-hildesheim.de

Christa Womser-Hacker
University of Hildesheim
Marienburger Platz 22
31141 Hildesheim
womser@uni-hildesheim.de

Werner Greve
University of Hildesheim
Marienburger Platz 22
31141 Hildesheim
wgreve@uni-hildesheim.de

ABSTRACT

Correlations between information retrieval system performance and user satisfaction are an important research topic. The expectation of users is a factor in most models of customer satisfaction in marketing research; however, it has not been used in experiments with information retrieval systems so far. In an interdisciplinary effort between information retrieval and psychology we developed an experimental design which uses the so-called confirmation/disconfirmation paradigm (C/D-paradigm) as a theoretical framework. This paradigm predicts that the satisfaction of users is strongly governed by their expectations towards a product or a system. We report a study with 89 participants in which two levels of both system performance and user expectation were tested. The results show that user expectation has an effect on the satisfaction as predicted by the C/D-paradigm. In addition, we confirmed previous studies which hint that system performance correlates with user satisfaction. The experiment also revealed that users significantly relax their relevance criteria and compensate for low system performance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Measurement, Human Factors

Keywords

Information retrieval systems, user studies, user satisfaction, user expectation, confirmation/disconfirmation paradigm

1. INTRODUCTION

Information retrieval systems provide their users with text documents for an information need that the user entered into the system. Retrieval systems are mostly evaluated with the so-called batch method. Independent jurors (assessors) judge whether documents are relevant for a search topic or not. This method is an aspect of the Cranfield paradigm [6, 13]. It makes objective results possible because the juror does not know from which system a result document comes. After jurors completed the relevance judgment process, it can be calculated how well systems perform according to measures like Mean Average Precision (MAP) or Binary Preference (BPref). Nevertheless, a problem of this approach is that the relevance assessment process cannot account for issues such as a user's knowledge and experience, situational relevance or user saturation [10].

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 setting, the performance of the users can be measured by observing the number of relevant documents they find. 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.

In recent years, there has been a growing concern on whether the results of batch and user experiments correlate. Do users have a better search experience, when systems improve in a batch comparison and contain more relevant documents in their result lists? Are changes in MAP noticeable for users in tests or real life? Are users more satisfied with better result lists and do better systems enable them to find more relevant documents? Some studies could not confirm this relation between system performance and user satisfaction, some did. Please note that, throughout this paper, we define the performance of a system as its ability to retrieve relevant documents according to a given query describing a users' information need. Previous research is presented in the following section.

In psychology, research on customer satisfaction has developed multi-dimensional models of satisfaction. These models take the expectation of the user into consideration. The C/D-paradigm predicts that the satisfaction judgment depends on both the product or system quality and the expectations of the users. Low expectations lead to satisfied users even for low system quality because the expectations of the user or customer are met. On the other hand, users with high expectations can only be satisfied if the system performs well. Otherwise their expectations are not confirmed and negative disconfirmation is likely [15]. An example may be a car enthusiast and a person who has no passion for cars. If these two types of customers are going to buy a new car, then the same car may meet the buyer's expectations in the one case, but fail in the other case. This simple example shows that the same stimulus can evoke completely different satisfaction responses depending on the underlying expectation. An interdisciplinary approach has great potential for user-based retrieval evaluation. The exchange of ideas between psychology and information retrieval has led to the study design described in this paper.

The study presented in this paper introduces expectation as one factor in a user based experiment. In addition to manipulate the system results to guarantee fixed performance measures, the test scenario manipulates the users to assume a certain expectation. The relation between expectations before using the system, the performance and the satisfaction after using the system are explored in the result section.

2. STATE OF THE ART

User-based retrieval evaluation focuses on the potential users of information retrieval systems. Järvelin and Ingwersen put the reason why the involvement of the user is so important for experimental retrieval evaluation in a nutshell: "The real issue in information retrieval systems design is not whether its recall-precision performance goes up by a statistically significant percentage. Rather, it is whether it helps the actor solve the search task more effectively or efficiently." [12] Recent studies can be classified into two types. On the one hand there are the satisfaction-oriented studies which concentrate on the relationship between system performance and user satisfaction. On the other hand there are the user performance-oriented studies which primarily analyze the influence of the system performance on the user performance.

2.1 Research on User Satisfaction

Compared to studies regarding the performance of users with retrieval systems, so far the satisfaction of users has been less frequently investigated in the context of interactive information retrieval. As can be seen from the following literature, due to the complexity of the evaluation task, a generalized method for assessing user satisfaction has not been established yet.

An investigation reported by Tagliacozzo [22] shows the difficulty of measuring user satisfaction. Therein Tagliacozzo analyzes the responses to a questionnaire sent to users of the MEDLINE bibliographic service. From the observation of inconsistencies in the users' judgment of the helpfulness and the usefulness of the MEDLINE search system, Tagliacozzo draws the conclusion that it is essential to tap several aspects of satisfaction in order to provide a more complete and accurate picture of the users satisfaction with the evaluated system [22].

Among other things, the issue of user satisfaction has been considered by Su [21], who proposed a user-oriented evaluation model for the information retrieval contexts. Besides various performance measures, subdivided into relevance, efficiency and connectivity measures, this model also includes several subjective measures of effectiveness, subdivided into utility and user satisfaction measures. Earlier research by Su [20] showed that the value or usefulness of search results as a whole seems to be a good predictor of information retrieval performance.

In 2006, Al-Maskari et al. [1] conducted a study in association with a submission to iCLEF2006. iCLEF is the interactive track of the Cross Language Evaluation Forum (CLEF[1]). In 2006, image retrieval was selected as the central theme of this track[2]. Pictures from the photo sharing community FLICKR[3] were used as data collection. Several effectiveness measures were used to evaluate the system performance (P@50norm; P@100norm; Q-measure; BPref-10; 10-Precision[4]). In addition, user performance was measured via task-specific modifications of recall and precision. Furthermore, 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 respectively satisfaction could be confirmed within the scope of this user study [1].

In another study, Al-Maskari et al. [2] conducted a similar experiment on the satisfaction of users. This time, participants directly searched the internet via the Google[5] search engine. The purpose of this study was to determine whether a correlation between the effectiveness of Google results quantified by Precision and Cumulative Gain measures (such as Cumulative Gain (CG), Discounted Cumulative Gain (DCG) and Normalized Discounted Cumulative Gain (NDCG)) 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. This user study could demonstrate a significant relationship between system performance and user satisfaction [2].

Huffman and Hochster [9] 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. Thus, Huffman and Hochster introduced a simple Cumulative Discounted Gain Measure which correlated with user satisfaction [9].

Also in 2007 Jansen et al. [11] studied the effect of branding on evaluation of system performance. In this experiment the same result documents were integrated in the search engine results pages of Google, MSN Live Search[6], Yahoo[7] and one unknown, in-house search engine. Participants completed four different search topics with supposedly four different search engines. In order to evaluate the users' perception of the system performance participants were asked to evaluate the result documents. The findings of this study show that branding has an effect on user's evaluation of system performance [11], which further implies that user expectations influence user satisfaction.

[1] http://www.clef-campaign.org

[2] http://nlp.uned.es/iCLEF/2006/guidelines.htm

[3] http://www.flickr.com/

[4] For a brief description of each measure see [1]

[5] http://www.google.co.uk

[6] Since Bing the new search engine by Microsoft started, the domain http://www.live.com forwards to the homepage from Bing http://www.bing.com/

[7] http://m.www.yahoo.com/

This short overview demonstrates that current research on user satisfaction is quite heterogeneous according to experimental setups and results. To obtain valid results, real-world search situations have to be simulated as closely as possible. For comparison between the results of different studies a standard measure for user satisfaction would be necessary. Furthermore, it can be noticed that user expectation as a variable has been almost neglected so far in user satisfaction research.

2.2 Research on User Performance

Evaluations regarding the performance of the user have become more common over the last years.

In the years 2000 and 2001 Turpin and Hersh [24] 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 [24].

The following two 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 [23] adopted MAP. The experiment of Allan et al. involved eight fixed system levels ranging from 0.5 to 0.98 BPref and the experiment of Turpin and Scholer involved five levels between 0.55 and 0.95 MAP. In addition, these two studies differ with respect to the underlying search tasks. Allan et al. chose a passage retrieval task in which the test subjects had to find, highlight and label all facets of the answer to a given information need. According to the BPref level they also differentiated between hard and easy topics. Turpin and Scholer found it especially important to use simple search tasks. Their precision-oriented task consisted in finding a document 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. Whereas the results of Allan et al. show statistically significant effects on the user performance at certain levels of BPref [4], the results of Turpin and Scholer only show a weak effect for the recall-oriented task [23].

Another interesting approach is reported by Al-Maskari et al. in 2008 [3]. They investigate the relationship between system performance and user performance as well as the system's influence on the user's satisfaction and perception of topic easiness. By using an application that facilitates the access to three different search engines it was possible to select the worst and the best system for every single topic. Thus system performance (measured in Average Precision (AP)) was specified on a per-topic basis. During the test, participants used both the superior and the inferior system without being aware of it. The user task was recall-oriented and consisted in identifying as many relevant documents as possible within a seven minute period of time. Interestingly, unlike in our present study, Al-Maskari et al. discovered significant correlations for a recall-oriented user performance measure. Besides the fact that participants of the superior system were able to find more relevant documents, results also show that they took less time and felt more satisfied [3].

Smith and Kantor also provide some interesting results on this issue [19]. Using a 3×3 factorial design they examined whether users alter their behavior in response to the performance of a search system. In the context of this study, subjects were told that they would be searching with the Google search engine. Similar to the study described just above, the result pages were selected on the fly. The test system would submit the original user queries to Google and than return result lists starting from different ranking positions depending on the actual system performance condition. Three different conditions were tested. The first condition was the standard system, which displayed result documents in the same order returned by Google. The remaining two conditions related to the two experimental systems, one condition with consistently-low-rankings (CLR) and one with inconsistently-low-rankings (ILR). In the CLR-condition the displayed lists constantly started at the 300th ranking position in the original Google result list. In the ILR-condition the starting positions within the original Google ranking were varied across one session for the first twenty documents of the displayed lists. The results obtained by Smith and Kantor strongly support our findings regarding the relaxation of the user-defined relevance criteria. As in the study presented in this paper, users of the bad system (the CLR-system) seemed to relax their relevance criteria and in the consequence thereof also accept documents of minor relevance [19].

Scholer and Turpin [17] 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 different search systems based on a four-level relevance scale of the documents. Findings suggest that different users adopt different relevance criteria. As a consequence, variances in system performance which can be observed in batch experiments are not necessarily visible in user experiments due to inhomogeneous relevance judgments of different users[17].

Most recently, Dostert and Kelly [8] conducted a study to investigate how and when users decide to end their search. The interactive information retrieval experiment was based on the TREC 2005 Robust Track collection. All subjects used the same test system and were asked to find relevant documents for four different search tasks. Besides recording the participants' actual recall, they were asked to estimate their achieved recall value and time spent searching. Results revealed that subjects recall estimates were not accurate, but positively correlated with actual recall. Furthermore, the time spent searching was also positively correlated with actual recall. A post-test interview showed that the main reason for terminating a search seems to be their assessment of enough information [8].

In summary, we may conclude that research results especially regarding the satisfaction of the user are not consistent and further study is needed to address the subject in a more detailed manner.

3. STUDY DESIGN

The objective of the present study was to develop a controlled study design which makes it possible to investigate the predictions of the C/D-paradigm in the information retrieval context. The research questions addressed were:

- What influence does the system performance have on user satisfaction and performance?

- What influence does the user expectation have on user satisfaction and performance?

To answer these questions our study adopts a between subjects design, in which each subject participates only in one treatment. That means that each participant is confronted with only one expectation manipulation. We assume that such a setup is more realistic and less prone to fail in achieving the objective.

In order to investigate the impact of system performance and expectations on user performance and satisfaction both of these independent variables had to be manipulated.

To control the system performance we used artificially constructed result lists for three different search topics, as it has been done in previous studies. For each of these topics two different result lists were created - one to simulate a low and one to simulate a high quality information retrieval system (subsequently also referred to as the inferior and the superior system or low and high system level). More details on the underlying system performance are presented in the subsection on the test system.

In order to test whether the users'satisfaction depends on their expectation, two different introduction sheets were designed. Therein the system was introduced as an expensive professional search system being about to be implemented in the university library in one case and as a student project with unknown quality in the other case. Test users read this description before the test and we expected that their expectation toward the system was affected. The introduction sheet also contained further information on the user test.

Since a between subjects design was adopted we randomly assigned the users to four treatment groups. In order to reach significant statements, we decided to test at least 20 subjects per group. Table 1 shows the two-by-two factorial design of our experiment. Our participants were not aware of the different experimental conditions.

		system performance	
		low	high
expectation	low	group1	group2
	high	group3	group4

Table 1: Two factor study design.

3.1 Topics and Documents

For the experiment, a standard IR evaluation collection was used in order to obtain a collection with topics and relevance assessments. We decided to use the German language corpus from the Cross Language Evaluation Forum (CLEF) which includes documents from the Swiss news agency *Schweizerische Depeschenagentur (SDA)*, the German daily newspaper *Frankfurter Rundschau (FR)* and the German weekly news magazine *Der Spiegel* [14]. The collection was used for a monolingual ad hoc retrieval setting.

From the large pool of topics available at CLEF, three were selected. Several criteria had to be met during the selection process. Finally, there needed to be a sufficient number of relevant documents available. Secondly, the topic should be about an issue which is still relevant and interesting for the test persons. This might not always be the case because the CLEF topics are developed for a test collection from the years 1993 and 1994.

The selected topics were the following ones:

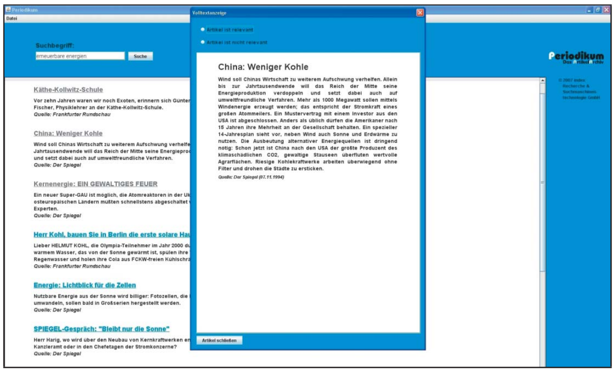
- C86: Renewable energy
- C187: Transports of nuclear material in Germany
- C190: Child labor in Asia

3.2 Subjects

Test users were recruited from our university. We selected only students who did not study any information technology related field. In order to avoid a gender effect, only female students were selected. Thus the results obtained from the study can only be generalized to the population of female searchers, but considering a gender-mixed sample would have either demanded a larger sample size or we had to live with the fact that a gender bias might affect the results. Users were motivated to participate in the study by an incentive. All users took part in a lottery for three cash prizes with a value of 100 Euro in total. Altogether, 89 participants completed the test which makes this experiment larger than most studies reported in the literature.

3.3 Test System

Design and functions of the application system resemble presently popular internet search engines in order to facilitate intuitive interaction. This was necessary to assure that interaction problems would not become a confounding variable. This was also necessary because no training session was given to the participants in order to shorten the duration of the test. A screenshot of the user interface can be seen in figure 1.



Käthe-Kollwitz-Schule

Vor zehn Jahren waren wir noch Exoten, erinnern sich Günter
Fischer, Physiklehrer an der Käthe-Kollwitz-Schule.
Quelle: Frankfurter Rundschau

China: Weniger Kohle

Wind soll Chinas Wirtschaft zu weiterem Aufschwung verhelfe
Jahrtausendwende will das Reich der Mitte seine Energieproc
und setzt dabei auch auf umweltfreundliche Verfahren.
Quelle: Der Spiegel

Kernenergie: EIN GEWALTIGES FEUER

Ein neuer Super-GAU ist möglich, die Atomreaktoren in der Uk
osteuropäischen Ländern mußten schnellstens abgeschaltet
Experten.
Quelle: Der Spiegel

Herr Kohl, bauen Sie in Berlin die erste solare Hau

Lieber HELMUT KOHL, die Olympia-Teilnehmer im Jahr 2000 du
warmem Wasser, das von der Sonne gewärmt ist, spülen ihre
Regenwasser und holen ihre Cola aus FCKW-freien Kühlschra
Quelle: Frankfurter Rundschau

Energie: Lichtblick für die Zellen

Nutzbare Energie aus der Sonne wird billiger: Fotozellen, die
umwandeln, sollen bald in Großserien hergestellt werden.
Quelle: Der Spiegel

SPIEGEL-Gespräch: "Bleibt nur die Sonne"

Herr Harig, wo wird über den Neubau von Kernkraftwerken en
Kanzleramt oder in den Chefetagen der Stromkonzerne?
Quelle: Der Spiegel

China: Weniger Kohle

Wind soll Chinas Wirtschaft zu weiterem Aufschwung verhelfen. Allein bis zur Jahrtausendwende will das Reich der Mitte seine Energieproduktion verdoppeln und setzt dabei auch auf umweltfreundliche Verfahren. Mehr als 1000 Megawatt sollen mittels Windenergie erzeugt werden; das entspricht der Stromkraft eines großen Atommeilers. Ein Mustervertrag mit einem Investor aus den USA ist abgeschlossen. Anders als üblich dürfen die Amerikaner nach 15 Jahren ihre Mehrheit an der Gesellschaft behalten. Ein spezieller 14-Jahresplan sieht vor, neben Wind auch Sonne und Erdwärme zu nutzen. Die Ausbeutung alternativer Energiequellen ist dringend nötig: Schon jetzt ist China nach den USA der größte Produzent des klimaschädlichen CO2, gewaltige Stauseen überfluten wertvolle Agrarflächen. Riesige Kohlekraftwerke arbeiten überwiegend ohne Filter und drohen die Städte zu ersticken.

Quelle: Der Spiegel (07.11.1994)

Figure 1: Screenshot of test application.

We adopted the algorithm presented by Turpin and Scholer for the construction of the result lists [23]. The algorithm starts with a random list and randomly swaps pairs of documents until the desired Average Precision is reached. The top positions are not treated specifically. It can be observed that result lists with the same Average Precision may look quite different. To avoid any issues we made sure that the top of the list mirrored the system level. In the first five documents, one or three irrelevant documents were placed for the high respectively low system level. Average Precision is defined as follows: "The mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved." [7]

Both Average Precision and Precision scores were selected within the range that has been used in previous studies (e.g. [23]).

For the superior system performance an Average Precision of 0.75 and for the inferior system performance an Average Precision of 0.55 was used.

Furthermore we varied the ratio of relevant to irrelevant documents (i.e. precision) for the high respectively low system level. For the superior system performance a precision of 0.6 and for the inferior system performance a precision of 0.5 was chosen. These values are within the range of previous experiments and represent the potential of modern search systems. For future experiments, these values should be empirically derived. Table 2 shows the ra-

		%	C86	C187	C190
documents available	relevant		60	57	50
	irrelevant		50	48	42
	total		110	105	92
inferior system	relevant	50%	50	48	42
	irrelevant	50%	50	48	42
	total		100	96	84
superior system	relevant	60%	60	57	50
	irrelevant	40%	40	39	34
	total		100	96	84

Table 2: Ratio of documents per topic.

tio of relevant and irrelevant documents for the three topics in our experiment.

In table 3 we provide the rankings of the six result lists used during the experiment. Each 0 represents an irrelevant document and each 1 an relevant document.

Each results page consisted of ten result documents. Results included the title of the document, snippet (first sentence of document) and document source. Our application system did not support keyword highlighting.

3.4 Experiment

The users came to a lab to conduct the test. After a welcome, they were given the written scenario for the test. The scenario was formulated to direct the test users toward the desired recall-oriented search process. The scenario was also used to create high and low expectations. Therefore, as already described in the context of the study design, one test group was told they would be using a professional search system, whereas the other group was told they would be using a search system, developed within a student project.

The search was motivated by the following scenario. All subjects were asked to imagine they were journalists and would use the presented search system to look for articles that concerned the topic of their next article. This scenario also fits the documents available within the CLEF collection. As previous work has shown such task-based scenarios help to minimize the artificiality of the test situation and at the same time motivate the participants [5, 18].

In the search tasks participants were required to find relevant documents according to the task-based scenario. To avoid confounding order effects the information needs were given to the users in different orders. Participants were given a maximum of ten minutes to complete each task, although they could terminate the search early if they felt they had completed the task. As in the case of the task-based scenario the opportunity to end the search early should help minimize the artificiality of the test situation and promote more natural behavior. Nevertheless, due to the study design, the typical real-world iterative search behavior was disabled. Queries were predetermined and users could not reformulate them in order to provide each subject with the same list of results. The explanation given to the participants to justify why the queries had to be predetermined was that we wanted to assure comparable conditions across participants.

The users could perform a limited number of actions in the user interface. They could browse through a list of result documents and evaluate them based on a representation by title and a snippet as in most commercial systems today. This approach was used to provide users with a somewhat familiar system design and at the same time maintain a high degree of control over the test situation for better comparison between users. Users could then select documents by clicking. The full text of the documents was shown in a separate window and users could then read them and decide if they were relevant or not. We want to point out that the subjects did not have to read through all documents in the result lists but only those that seemed relevant to them based on title and snippet. Once selected by a user, the user had to decide whether or not the document was relevant. The judgments as well as timestamps for all interactions with the application system were recorded in a log file. Oral comments of the users were not systematically recorded by the test conductor.

Finally, after finishing the search tasks for all three topics, each subject completed a questionnaire regarding the overall satisfaction, that is to say, there was not a single questionnaire after each search task. The questionnaire contained 28 question items and dealt with both ease of use of the application system and satisfaction with the presented result lists. The questionnaire also included a section on personal data, which inquired about age, occupation, years of computer and internet use. In terms of content the satisfaction part of the questionnaire includes questions regarding satisfaction with ease of use, efficiency, output display, precision, ranking of results, result quality and reuse probability. Therefore, the questionnaire covers similar content areas to the studies described in section 2.1. Most questions were rated on a 7-point scale which ranged from 1 (completely correct) to 7 (not at all correct). We also tried to indirectly determine the satisfaction of the test users with the search results. Therefore we asked them at the end of the test if they were willing to participate in another user test. The underlying assumption was that it is more likely that satisfied subjects would agree to assist in a second test.

4. RESULTS AND DISCUSSION

The results of the experiments were analyzed according to the validity of the test design and subsequently the (subjective) user perception and the (objective) user performance were analyzed.

4.1 Test Design

Prior to user based analysis, several validation checks were performed to ensure that there were no confounding variables that had an additional effect on the collected data. We briefly mention two effects that are connected with the search tasks.

Similar to the results of Turpin and Scholer [23] the one-way ANOVA analysis of our data indicates that there is a statistically significant topic effect. The recorded user performance for the three search tasks was included as repeated measurement factor. Our observations suggest that comparatively topic C86 was more difficult and C187 more easy to deal with. We assume that the existence of such a topic effect can be looked upon favorably because different information needs in realistic search situations also vary in their level of difficulty.

As already mentioned the information needs were given to the test users in different orders. In this manner we intended to control possible order effects such as learning and fatigue. The relationship between user performance and topic order were examined using a one-way ANOVA, with the order as independent factor. Only for the topics C86 and C190 significant order effects can be observed. This supports the above observation that topic C187 was in general easier to deal with and therefore no training was necessary for the test subjects. This check also shows that the variation of the topic order was appropriate.

In most cases, subjects did not terminate their search early (on average 10 participants (SD = 1.7) searched less than 10 minutes per task). In contrast to the results obtained by Dostert and Kelly [8], in our study spending more time searching did not lead to better recall. However, early search termination did correlate with the

topic	system level	ranking
C86	low	1 1 0 0 0 1 0 1 0 0 1 1 1 0 0 0 1 1 0 1 1 1 1 1 0 0 0 0 1 1 1 0 0 1 1 1 0 1 1 0 1 0 1 1 0 1 1 0 1 0 1 0 1 1 0 0 0 1 0 0 0 0 1 1 1 0 1 0 0 1 0 0 1 1 0 0 1 0 0 0 0 1 1 0 1 1 1 0 0 0 1 0 0 1 1 1 0 1 0 0
	high	1 1 0 1 1 1 1 0 0 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 0 1 0 1 0 0 0 0 0 1 1 1 0 1 0 0 1 1 1 0 1 1 1 1 1 0 0 1 1 0 1 0 1 1 1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 1 1 1 10 0 0 0 0 0 0 1 0
C187	low	0 1 1 0 0 0 1 1 1 0 1 1 1 1 0 1 0 1 1 1 0 0 1 0 0 0 1 0 0 1 1 1 0 0 1 0 0 1 0 0 0 1 1 1 0 1 1 1 1 1 0 0 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0 1 0 0 0 1 1 0 0 1 0
	high	0 1 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 0 1 1 0 1 1 1 0 0 0 1 1 1 1 1 0 1 0 1 1 1 0 0 0 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 10 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0
C190	low	1 1 0 0 0 1 0 1 1 1 0 1 0 0 1 1 0 1 0 1 1 1 0 1 1 0 0 0 0 1 1 0 0 0 1 0 0 0 0 1 0 0 1 1 0 1 1 0 1 1 1 0 1 1 1 0 0 1 1 0 1 1 0 0 1 1 1 0 1 0 0 0 0 0 0 1 1 0 0 1 1 1 0 0 0
	high	1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 1 0 1 0 1 0 1 0 1 0 0 1 0 1 0 1 0 0 1 0 1 1 1 0 0 0 1 0 1 0 1 1 0 1 1 1 1 0 0 1 1 0 1 1 0 0 0 1 0 0 0 1 1 0 0 0 1 1

Table 3: Ranking of result lists.

time taken to find the first relevant document for two of the three topics (C86/C187 p = 0.0). Thus it seems that some participants searched slightly faster than others. As this behavior reflects real-life searching and those participants are equally distributed across the four treatment groups, we assume that it adds to the potential generalization of our findings.

The influence of expectations and system performance on user satisfaction and user performance was for the most part examined using a two-way ANOVA. No significant influence could be detected for the expectation of the users. One possible interpretation could be that expectations do not have an influence on the perception of retrieval results. Far more likely however is the interpretation that the manipulation was not sufficient. This assumption was confirmed by some subjects within informal conversations at the end of the test. These observations suggest that further research is needed to clarify the validity of the confirmation/disconfirmation paradigm in the field of information retrieval.

With respect to the system performance on the contrary significant influences could be identified for both user satisfaction and user performance. These group differences are described in the following sections.

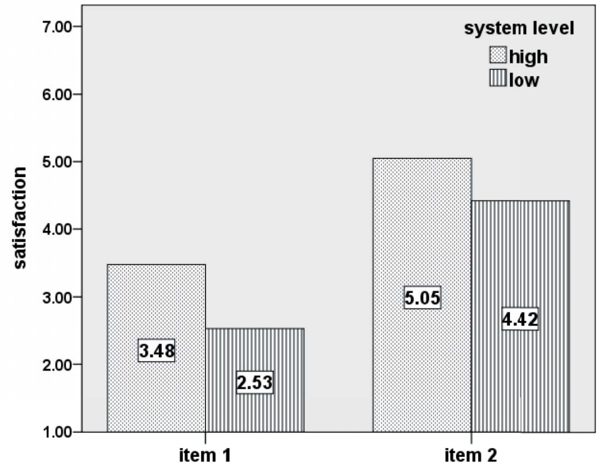


Figure 2: Satisfaction precision-oriented question items.

4.2 User Satisfaction

At first, we separately analyzed the question items using a two-way ANOVA. Corresponding to the underlying study design, the system performance and the user expectation form the independent variables. The individual responses were included as dependent variable. Significant group differences in relation to the system level only occurred for the following precision-oriented items:

- Item 1: The filtering of articles could have been better. (p = 0.008)
- Item 2: Most articles have been relevant with respect of the queries. (p = 0.025)

Figure 2 shows the results for these two questions. It must be pointed out that the first question is formulated negatively. For better comparison, we inverted the scale for item 2, so that now for both questions higher values correspond to higher satisfaction. In figure 2 it can be seen that in both cases subjects that used the superior system were also more satisfied with the presented performance of the system. This shows that users are indeed able to perceive improvements in system performance. Nevertheless, even though almost no significant group differences have occurred in this first analysis, in the next step we tried to combine several items to one scale. Therefore we only used those items that dealt with the user's satisfaction with the presented result lists. An analysis of reliability has been adopted to test the quality of this scale and to choose the most appropriate items to be included. We applied Cronbach's Alpha as a measure of internal test reliability to identify a valid group of items from our questionnaire. The best Cronbach's Alpha of 0.69 (which is close to the as sufficient regarded value of 0.7) is achieved by combining the following items:

- Item 1: The filtering of articles could have been better.
- Item 2: Most articles have been relevant with respect of the queries.
- Item 3: I am satisfied with the quality of the search results.
- Item 4: The presentation of the results was clearly structured.
- Item 5: The order of the search results reflected the relevance of the articles.
- Item 6: The articles I selected were helpful for the search.

The two-way ANOVA of the combined scale reinforces the results from the individual analysis, as can be seen in figure 3.

The predictions of the C/D-paradigm can be observed in our experimental data. Nevertheless, we did not find statistically significant differences for the user expectations (p = 0.50). However, as already shown for question items 1 and 2 there is a significant difference (p = 0.01) between the two system levels. There occurred no interaction effects (p = 0.78).

Figure 3 plots the mean values for the four test groups based on the combined satisfaction scale specified above. Of the two groups that used the inferior system the group with the treatment - high expectation - is less satisfied (4.15 vs. 4.32) because its expectations were not fulfilled (negative disconfirmation). Furthermore, of the two groups that used the superior system the group with the treatment - low expectation - is more satisfied (4.65 vs. 4.72) because its expectations were exceeded (positive disconfirmation).

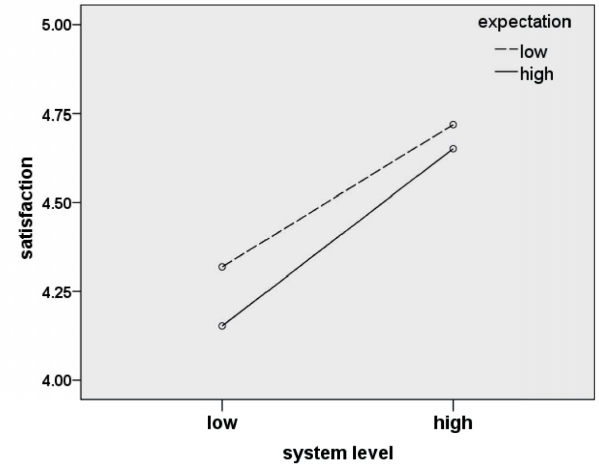


Figure 3: Confirmation/disconfirmation paradigm.

The indirect satisfaction test with the additional user test did not lead to conclusive results. Altogether 76 of the 89 subjects offered to participate in another test. Such a high percentage (85%) suggests that this result does not represent the indirect satisfaction of the users. Instead, it seems that they felt obliged to offer their willingness to attend a further test.

The fact that correlations between system performance and user satisfaction are rarely observed also reflects the fact that experimental variables for such studies are hard to evaluate. As Al-Maskari et al. [3] already pointed out the satisfaction of the user "[...] is always likely to exhibit discrepancies amongst individuals."

4.3 User Performance

In addition to the satisfaction as expressed by the test users, we also intended to record and analyze their performance. Five user performance measures were used to investigate the experimental data:

- Doc@10: Number of relevant documents that users were able to identify within the given time of ten minutes per task. In the experiment of Turpin and Scholer subjects had a five minute time limit per topic [23].

- User Recall (UR): Number of documents correctly identified as relevant divided by total number of relevant documents in the result list. Al-Maskari et al. pursued a similar strategy within the framework of their user study. But since this study dealt with image retrieval the focus was on identifying unique documents respectively images [1].

- $t_{1.Doc}$: Following Turpin and Scholer we compared the time that users took to find their first relevant document [23].

- User Precision (UP): According to Al-Maskari et al. the UP was calculated as number of documents correctly identified as relevant divided by total number of documents judged as relevant by the user [1].

- Pre-Click-Precision (PCP): Number of documents correctly identified as relevant divided by total number of documents selected by the user to read the full text. Thus PCP catches the surface impression that the users had from the result lists. This performance measure was derived from the idea of Resnick and Lergier to ask participants about their "[...] expectation of how well his/her selection would match his/her expectations." [16]. In this context they introduced the term *Pre-click confidence*.

In the focus of the analysis of the user performance was the mean performance of the users over all three tasks and not per individual task. Thus we are able to make more general and more widely applicable statements.

Again a two-way ANOVA was used to determine whether significant differences existed between the four test groups. The five user performance measures were one after another included as dependent variables. Table 4 summarizes the significance values regarding the main and interaction effects of the independent factors with one degree of freedom.

	expectation		system performance		interaction	
measure	F[1]	p[2]	F	p	F	p
Doc@10	2.868	0.094	0.465	0.497	2.731	0.102
UR	2.562	0.113	0.519	0.473	2.316	0.132
$t_{1.Doc}$	0.041	0.84	0.101	0.751	0.289	0.592
UP	0.486	0.488	13.045	0.001	4.823	0.031
PCP	0.939	0.335	4.424	0.038	1.013	0.317

[1] F-value
[2] Significance value

Table 4: Summary ANOVA user performance.

These results show that the system performance has a significant impact solely on the precision-oriented measures, UP and PCP. For the remaining measures none of the test conditions has a significant influence (p > 0.05). The fact that there are no significant main effects for the recall-oriented measures, Doc@10 and UR, indicates that users are able to compensate the difference between the two system levels for these performance measures.

The significant average value difference for PCP between the high and low system performance group can be explained as a reinforcement effect. Although the results for the Doc@10 are not significant (cf. Table 4) users of the inferior system by trend found fewer relevant documents (8.75) than users of the superior system (9.55). In addition a weak, but also not significant trend could be observed that users with low system level selected more documents to read the full text (13.87) than users with high system level (13.49). In themselves the differences are not significant but for the calculation of PCP these two quantities are divided by one another. Thus the non significant differences are reinforced by the user performance measure.

The most interesting result concerning the user performance is the significant difference in the User Precision for the two system levels. Figure 4 illustrates this difference. In the case of using the superior system subjects on average achieved a higher Precision value (0.93) than in the case of using the inferior system (0.86). The percentage difference is 8%. At first it might be surprising to see that the test users were not able to compensate the difference in system performance for the User Precision as well. Figure 5 shows why the detected differences occurred. The y-axis is sorted by the number of documents. As the figure shows, subjects who were using the inferior system erroneously judged more documents as relevant than users of the superior system did. In addition, the same subjects judged fewer documents erroneously as irrelevant compared to users of the superior system. Erroneously refers to the relevance judgment provided by the CLEF jurors which was considered as ground truth for our experiment. At this point we also want to remark that non-relevant documents in our result lists correspond to explicitly CLEF-judged non-relevant documents.

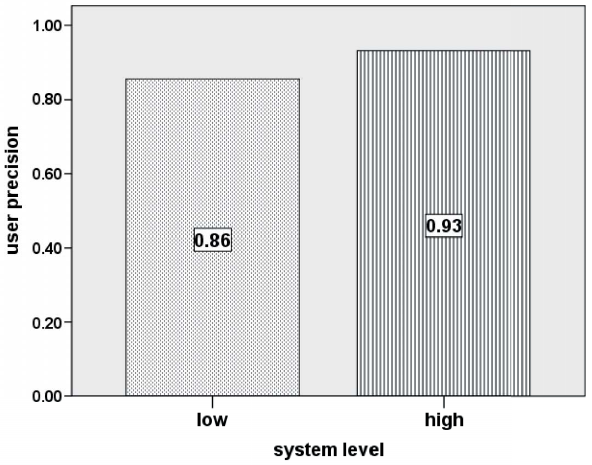


Figure 4: User precision.

These findings suggest that users assimilate their relevance criteria to some extent to the quality of the information retrieval system. For result lists with a wide range of relevant documents, users seem to be stricter in their relevance judgment than for result lists with few relevant documents. This type of adaptive search behavior has also been observed in previous user studies with different experimental setups [19, 17].

The difference for the User Precision is 8% and cannot be considered as large. Nevertheless, the percentage differences for the number of incorrectly judged documents are larger. For the documents incorrectly judged as relevant a difference of 57% (inferior system: 1.68 vs. superior system: 0.73) was perceived. For the documents incorrectly judged as irrelevant, 28% (inferior system: 1.64 vs. superior system: 2.29) were measured. These numbers can be compared to interrater agreement experiments carried out with TREC data. When more than one juror judged the documents for one topic, they agreed only on 30% to 40% of the relevant documents [25]. To draw the conclusion, the perception of relevance seems to be influenced by the context, in the case of our experiment, the system performance level.

4.4 Analysis of Covariance

In order to account for different levels of experience in using information retrieval systems we ran an analysis of covariance with the user's search experience as covariate. For this purpose the questionnaire contained items that recorded how many hours per week the test users spent on average with computer work, online etc.

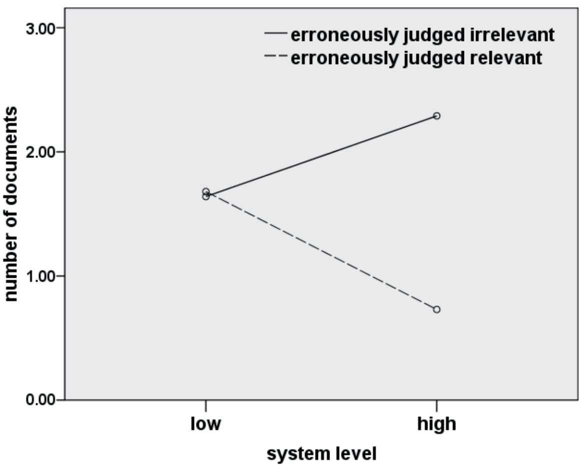


Figure 5: Assimilation effect.

These items were used to classify users as superior versus average experienced users. The search engine experience reported by the test users was normally distributed. However, the search experience neither had a significant impact on the user satisfaction nor the user performance.

Our experiment shows that user reported relevance as it is used in many interactive IR studies is subject to many contextual factors which may not be controlled in each experiment. We show that the system performance is one context factor which has an influence. As a consequence, user studies should not solely rely on user reported relevance but introduce some sort of objective assessment. The same is true for click-through data.

4.5 Conclusions

The primary aim of our experiment was to investigate the influence that the users' expectations about the performance of a search system have on their satisfaction and performance with retrieval systems. The experimental setup is motivated by the confirmation/disconfirmation paradigm, a widely used model to describe customer satisfaction in marketing research. Besides studying the influence of user expectations, we also investigated the influence that the system performance has on both dependent variables. Although this study only constitutes a first attempt to apply the concept of expectation confirmation to interactive retrieval evaluation, we already found some indication that the users'expectation of system performance may influence their satisfaction as predicted by the confirmation/disconfirmation paradigm. Nevertheless, the differences were small and not significant according to the statistical analysis undertaken. The same is true for relation between the subjects' expectation and performance, the expectation manipulation revealed no significant differences in the user performance measures between the four experimental groups. With respect to the system performance, however, we found that self-reported relevance in user studies is highly context dependent. This can be seen by the fact that users significantly seem to relax their relevance criteria as soon as they begin to use the lower quality search engine. We also found that the subjects' statements of satisfaction correlated positively with the actual system performance level.

5. OUTLOOK

In future studies, we intend to further elaborate the concept of customer expectation in the context of information retrieval. The expectation can be manipulated stronger to see if there is a larger and significant effect on the satisfaction and performance of the user. Another strategy to overcome the weaknesses attributed to the operationalisation of the expectation manipulation may be to let each participant compare two treatments.

It is also interesting and important for future research to allow participants to use individual search strategies and in this context further investigate the observed assimilation effect. For user studies, between subjects test design should be preferred or at least combined with within subjects setups. Another important aspect for future experiments is the degree to which users relax their relevance criteria. It should be measured in respect to the system performance and other factors.

The results of our study show that self reported relevance in user studies is highly context dependent. The relevance criteria are affected by the system performance level. As a consequence, past experiments without a ground truth for relevance need to be interpreted carefully.

6. REFERENCES

[1] A. Al-Maskari, P. Clough, and M. Sanderson. Users' effectiveness and satisfaction for image retrieval. In *Proc. LWA '06*, pages 84–88, 2006. http://web1.bib.uni-hildesheim.de/edocs/2007/522070116/doc/522070116.pdf.

[2] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *Proc. SIGIR '07*, pages 773–774. ACM Press, 2007.

[3] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: does the test collection predict users' effectiveness? In *Proc. SIGIR '08*, pages 59–66. ACM Press, 2008.

[4] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proc. SIGIR '05*, pages 433–440. ACM Press, 2005.

[5] P. Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1):71–90, 2000.

[6] C. Buckley and E. Voorhees. Retrieval system evaluation. In *TREC: Experiment and Evaluation in Information Retrieval*, pages 53–75. Cambridge & London: MIT Press, 2005.

[7] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. SIGIR '00*, pages 33–40. ACM Press, 2000.

[8] M. Dostert and D. Kelly. Users' stopping behaviors and estimates of recall. In *Proc. SIGIR '09*, pages 820–821. ACM Press, 2009.

[9] S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *Proc. SIGIR '07*, pages 567–574. ACM Press, 2007.

[10] P. Ingwersen and K. Järvelin. *The turn: integration of information seeking and retrieval in context*. Springer, Dordrecht, 2005.

[11] B. J. Jansen, M. Zhang, and Y. Zhang. The effect of brand awareness on the evaluation of search engine results. In *Proc. CHI '07*, pages 2471–2476. ACM Press, 2007.

[12] K. Järvelin and P. Ingwersen. Information seeking research needs extension toward tasks and technology. *Information Research*, 10(1), 2004. http://informationr.net/ir/10-1/paper212.html.

[13] T. Mandl. Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance. *Informatica*, 32:27Ű–38, 2008.

[14] G. M. Nunzio, N. Ferro, T. Mandl, and C. Peters. Clef 2007: Ad hoc track overview. In *Proc. CLEF '07*, pages 13–32. Springer, 2008.

[15] P. G. Patterson. Expectations and product performance as determinants of satisfaction for a high-involvement purchase. *Psychology and Marketing*, 10(5):449–465, 1993.

[16] M. L. Resnick and R. Lergier. Task specific user strategies in on-line search. *Journal of E-Business*, 3(1):1–22, 2003.

[17] F. Scholer and A. Turpin. Relevance thresholds in system evaluations. In *Proc. SIGIR '08*, pages 693–694. ACM Press, 2008.

[18] F. Scholer, A. H. Turpin, and M. Wu. Measuring user relevance criteria. In *Proc. EVIA '08*, pages 47–56, 2008. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/EVIA2008/09-EVIA2008-ScholerF.pdf.

[19] C. L. Smith and P. B. Kantor. User adaptation: Good results from poor systems. In *Proc. SIGIR '08*, pages 147–154. ACM Press, 2008.

[20] L. T. Su. Value of search results as a whole as the best single measure of information retrieval performance. *Information Processing & Management*, 34(5):557–579, 1998.

[21] L. T. Su. A comprehensive and systematic model of user evaluation of web search engines: I. theory and background. *JASIST*, 54(13):1175–1192, 2003.

[22] R. Tagliacozzo. Estimating the satisfaction of information users. *Bull. Med. Libr. Assoc.*, 65(2):243–249, 1977.

[23] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. SIGIR '06*, pages 11–18. ACM Press, 2006.

[24] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proc. SIGIR '01*, pages 225–231. ACM Press, 2001.

[25] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.