A Review of Factors Influencing User Satisfaction in Information Retrieval

Azzah Al-Maskari
Directorate General of Technological Education
P.O.Box: 413, Post code 100
Tel: +96824839124
Fax: +96824813723
Email: maskeri.a@gmail.com

Mark Sanderson
Department of Information Studies, University of Sheffield
Regent Court, 211 Portobello Street
Sheffield S1 4DP, UK
Tel: +44 114 22 22648
Fax: +44 114 27 80300
Email: m.sanderson@shef.ac.uk

A Review of Factors Influencing User Satisfaction in Information Retrieval

Abstract

This paper investigates factors influencing user satisfaction in information retrieval. It is evident from this study that user satisfaction is a subjective variable which can be influenced by several factors such as system effectiveness, user effectiveness, user effort and user characteristics and expectations. Therefore, information retrieval evaluators should consider all these factors in obtaining user satisfaction and in using it as a criterion of system effectiveness. Previous studies have conflicting conclusion on the relationship between user satisfaction and system effectiveness, this study has substantiated this relationship and supports using user satisfaction as a criterion of system effectiveness.

1. Introduction

The main aim of an information retrieval (IR) system is to satisfy the need of its users. Lancaster (1979; 1981) suggests that an IR system can be evaluated according to three criteria: (i) the suitability of a system in terms of the specific IR tasks for which it will be used; (ii) the system's task performance efficiency and (iii) the extent to which the system satisfies the information needs of its users. Therefore user satisfaction is an important factor in evaluating IR systems. User satisfaction in IR research is generally considered a criterion of system success and effectiveness. Griffiths et al. (2007) surveyed the information retrieval and information system (IS) literature in an attempt to understand what constitutes user satisfaction and the factors affect it. They found that user satisfaction is a measure that has been considered immensely in user-oriented system evaluation within both the IR and IS literature. According to their survey they found that user satisfaction is not a single construct upon which to base user assessment of system effectiveness but is influenced by other factors, including: system output; user expectation and attitude, perceived ease of use and usefulness; system type; and task difficulty.

While there is much research calling for the importance of user satisfaction as a criterion of IR system evaluation (e.g. Spärck Jones, 1981; Gatian, 1994; Gluck, 1996; Huffman and Hochster, 2007) there is no comprehensive study investigating factors influencing user satisfaction. For example, several studies examined the relationship between user satisfaction and system effectiveness[1] (e.g. Huffman and Hochster, 2007; Thomas and Hawking, 2006; Johnson et al., 2003; Turpin and Hersh, 2001; Sandore, 1990; Gluck, 1996); however, these studies did not consider user experience and expectation of the IR system under evaluation. This paper discusses results that emerged from an experiment that was designed to investigate the factors influencing user satisfaction in the IR field. This paper first defines user satisfaction (section 2) and then defines the factors influencing it: system effectiveness (section 3.1), user effectiveness (section 3.2), user effort (section 3.3) and user characteristics (section 3.4). Section 4 presents past

[1] As quantified by the IR effectiveness measures such as precision and mean average precision.

work and list of hypotheses tested in this paper. Sections 5 presents results of the experiment conducted to examine the factors influencing user satisfaction; section 6 and 7 present a discussion and the conclusions of this work.

2. Definition of user satisfaction and related research

In an early attempt to define 'user satisfaction' as a concept, Tessier et al. (1977) stated that satisfaction was "ultimately a state experienced inside the user's head" (p.383) and therefore was a response that "may be both intellectual and emotional" (p.384).

Spärck Jones (1981: p.55) stressed the importance of 'user satisfaction' and considered it as the basic concept of information retrieval system evaluation that could not be ignored in any experiment. User satisfaction has the following advantages, as identified by Su (1992): (i) it takes explicit account of users and their subjective evaluation of various aspects of the IR interaction; (ii) it focuses on multi-dimensional evaluation of the interactive processes and (iii) it also recognizes user and request characteristics as among possible influencing factors in user evaluation.

In 1973, Cooper described "utility" which required users to indicate their satisfaction with search results by assigning a monetary value to the retrieved documents. Soergel (1976) rejected Cooper's proposal that user satisfaction with search results was a valid measure of retrieval. Soergel discarded user satisfaction as a measure, maintaining that users may be satisfied with less than optimal search results, especially if a definitive assessment is made only for the first iteration of results returned by the system. In particular, he cited the "user-distraction" phenomenon whereby a user, upon receiving an irrelevant document from the IR system in response to some search operation, might still express satisfaction with the irrelevant search result. He recommended, therefore, that helping users in completing their search tasks successfully should take priority over seeking their satisfaction.

Belkin and Vickery (1985) warned, like Tessier et al. before them, of the many problems associated with satisfaction criteria. These problems arose from the ambiguous definition of 'satisfaction' and how to measure it. Hildreth (2001) further questioned the reliability of the satisfaction criterion as a measure due to its lack of independence from other influential factors in the retrieval procedure. When used as a performance measure in IR system evaluation, it can be easily affected by non-performance factors that may confound the results. This concern was especially critical if the actual performance factors being measured were the quality of search results or assessments of search success as judged by the users. Hildreth posited that end users of IR systems often expressed satisfaction both with their results and with the overall performance of the system, even when objective analysis of the results showed them to be poor. Moreover, Hildreth argued that evaluation studies that relied on measures such as user perception of ease of use and subjective satisfaction with the search results did not provide a clear and consistent answer as to how user satisfaction may predict their actual search effectiveness. He found that user perception of ease of use had an effect, possibly greater than the results themselves, on user satisfaction.

Harter and Hert (1997) reported that satisfaction has been the most widely used evaluation concept in information system evaluation. The authors reviewed the literature on Management of Information Systems (MIS) and Library Information Systems (LIS) on the use of the satisfaction criterion in information system research and evaluation.

Previous studies had substantially different conclusions on the applicability of user satisfaction in user-oriented evaluation. For example, Hildreth (2001) asserted that user satisfaction was a false measure when used in

predicting system success; while other researchers (e.g., Gatian, 1994; Gluck, 1996; Huffman and Hochster, 2007) found that user satisfaction was significantly associated with system effectiveness. According to previous studies there is a confusing picture on the applicability of user satisfaction as a measure of system effectiveness. Therefore, this paper clarifies this doubt and provides a clearer picture on the relationship between user satisfaction and system effectiveness. The results of experiments showed a significant correlation between user satisfaction and system effectiveness. Furthermore, it was illustrated that while users searching in two systems with different effectiveness; users were significantly more satisfied with the superior system as compared to their satisfaction with the inferior system. Results from this study also confirmed that user satisfaction was influenced by several factors such as, user effectiveness, system effectiveness, user effort, and user expectation.

3. Factors influencing user satisfaction

In this paper, we investigate the relationship between user satisfaction and the following four factors: system effectiveness, user effectiveness, user effort, and user characteristics. These factors are explained in the following paragraphs.

3.1 System effectiveness

System effectiveness measures how well a given IR system achieves its objective. Traditionally, system retrieval effectiveness is measured in terms of precision (the fraction of retrieved documents retrieved by the IR system that are also relevant to the query) and recall (the fraction of the relevant documents present in the database that are retrieved by the IR system). These two parameters characterise the ability of the system to retrieve relevant documents and avoid irrelevant ones (Van Rijsbergen, 1979: p.114). Other effectiveness measures are discussed in Korfhage, 1997; Baeza-Yates and Ribeiro-Neto, 1999; and Järvelin and Kekäläinen, 2000.

3.2 User effectiveness

User effectiveness is defined as the accuracy and completeness with which users achieve certain goals. User effectiveness can be measured by the following criteria: (i) the number of tasks successfully completed, (ii) number of relevant documents obtained, and (iii) the time taken by users to complete set tasks (Hersh, et al., 2000; Turpin and Hersh, 2001; Allan et al., 2005; Turpin and Scholer, 2006; Frøkjær et al., 2000; Lazonder et al., 2000). Indicators of effectiveness also include quality of solution and error rates. User effectiveness is different from system effectiveness, for example system effectiveness is measured objectively by the number of relevant documents retrieved by the IR system (e.g. TREC relevance assessments) whereas user effectiveness is measured by the number of relevant documents saved by the users from the number of relevant documents retrieved by the IR system (e.g. the number of relevant documents identified by the users and at the same time match with TREC relevance assessments).

3.3 User effort

User effort can be defined in a similar way to the definition of ‘information searching behaviour’ (Wilson, 2000); information searching behaviour is the user search behaviour when interacting with an IR system to search for

relevant information. User effort can be measured by the number of clicks, number of queries and number of query reformulations, and rank position accessed to obtain relevant information.

Lancaster (1969) also considers the amount of effort expended during the search as one the critical features affecting user satisfaction. Expected search length (ESL) by Cooper (1968) is also a form of user effort; ESL is the average number of documents examined to retrieve a given number of relevant documents.

3.4 User characteristics

Human factors and individual differences were recognized as a key aspect for understanding users search behaviours (Egan, 1988; Nielsen, 1993). Some of these factors were closely related to user characteristics/traits such as familiarity with the search topic (domain expertise), motivation, and experience in various aspects such as computing, librarianship, and skills in searching for information. Marchionini (1995) explained that every individual has a unique set of IR skills. These skills consist of three components: (i) domain expertise, (ii) system expertise and (iii) search expertise. Therefore, the term "user characteristics" – as used in this paper – are synonymous with Marchioni's definitions, i.e. users' familiarity with the search topics, and their search experience.

4 Hypotheses Tested

In this paper, we examine four hypotheses (Figure 1) which investigated the relationship between user satisfaction and the four factors explained in section 3: system effectiveness, user effectiveness, user effort, and user characteristics. This section summarises the results from previous research related to the influence of each factor on user satisfaction.

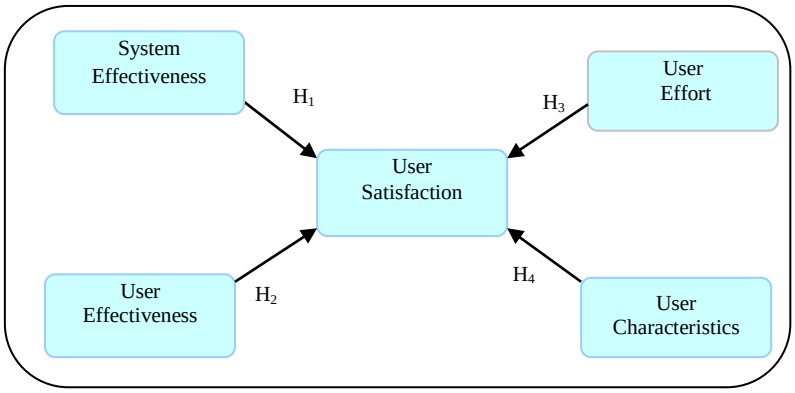

Figure 1: *Factors and hypotheses tested*

4.1 H1— System effectiveness influences user satisfaction

It is expected that system effectiveness is correlated with user satisfaction. The relationship between system effectiveness and user satisfaction was investigated by several researchers. Huffman and Hochster (2007) observed a strong correlation between the relevance of results and user satisfaction using navigational and non-navigational queries. In their investigation, seven participants assessed the relevance of the first three results of a list of documents returned by Google for 200 queries and user satisfaction ratings for the results. They found that the

relationship between relevance and satisfaction weakened rapidly after the first position for navigational queries, while it stays constantly the same strength at all three positions for non-navigational queries.

Thomas and Hawking (2006) presented 23 users with two side-by-side set of results of high and low quality: the high-quality screen displayed the first ten results of Google and the low-quality screen displayed the results from 21-30. Users indicated their preference of the two sets of the results and they successfully distinguished between the high quality and the low quality results. Johnson et al.(2003) recruited 23 participants to search on three engines (Excite, NorthernLight, and HotBot) for their own information need. They observed a strong correlation between user's satisfaction with precision of the results and judgement of the systems' effectiveness (system effectiveness was measured according to the degree of relevance of the items retrieved as rated by the users).

However, Turpin and Hersh (2001) did not substantiate a relationship between system effectiveness and user satisfaction. Twenty-four users were involved and required to identify a number of factual answers to eight questions from two systems with different effectiveness with MAP[2] scores of 0.27 and 0.35. Despite the systems exhibiting quite different retrieval effectiveness, there was no significant difference in user satisfaction with the results retrieved from the systems.

Saracevic and Kantor (1988) after their extensive study of on-line database searchers found that "satisfaction with results" correlated with precision but not recall of the search results. However, different findings relating to the relationship between user satisfaction and precision and recall were reported by Su (1992; 1992; 1994; 1998) who found that the user satisfaction with the completeness of the results correlated higher with their judgment of system success[3] than their satisfaction with precision of the results. Sandore (1990) also reported finding a low correlation between precision and satisfaction; users were often satisfied with low precision search results, even in cases where their goal was to achieve high precision results. The reason Su found recall to be more important than precision in evaluating the IR system success may be attributable to the users' professional status and users' purpose of the search (the majority of the participants in Su's study were PhD students and academic faculty members with a need to obtain information for writing up dissertation or grant proposals). In a follow-up study, Su (2003) reported results which contradicted her previous investigation; the (36) users preferred precision over recall when conducting their search tasks using four search engines (Alta Vista, Excite, Infoseek and Lycos). In the latter study, the purpose of the search was different from that which obtained for Su's initial studies; participants were undergraduates searching for the purpose of class assignments, personal interests, graduates schools, travel and jobs.

Hufnagel (1990) also questioned the validity of employing user satisfaction ratings as a measure of system effectiveness. He argued that user satisfaction ratings may be a reflection of individual performance outcomes (i.e. success or failure), rather than an objective assessment of overall system effectiveness. In Hufnagel's (1990) study, eighty MBA students participated in a laboratory study for the purpose of evaluating eight accounting computer systems; these students were asked to solve a series of standard accounting problems using the tested systems that

[2] Mean Average Effectiveness, it is average precision obtained after each relevant document is retrieved

[3] User's judgment of system success in providing help for information needs or problems.

had been covered in an MBA course. Students were asked to indicate the extent to which they believed their performance was affected by a variety of different factors, including the amount of effort expended, the quality of the computer system used, how well they understood the system, any unanticipated factors significantly influencing the outcome ("good luck/bad luck"), and the difficulty of the task itself. Results indicated that those users who successfully performed the task attributed their performance outcomes to their own effort and understanding, while those who were unsuccessful tended to blame their poor performance on luck and/or the quality of the system. Hufnagel (1990) concluded that the actual contribution of the system is ambiguous and difficult to quantify from the users' perspective, because users tend to discount the contribution of the computer system when things go well and to blame the system when things go poorly; thus, Hufnagel suggested that user satisfaction is not an adequate measure for system effectiveness.

Gluck (1996) provided a complimentary review of the major research on user satisfaction that has appeared in the LIS and MIS literature. Gluck (1996) reported a strong correlation between user satisfaction with retrieved items and the relevance of these items.

4.2 H_2 — User effectiveness influences user satisfaction

It is expected that user effectiveness (as measured by the number of relevant documents identified and/or the time taken users to complete the task) correlates with user satisfaction: as user effectiveness decreases, user satisfaction will correspondingly decrease. Su (2003) and Law et al. (2006) both concluded that user satisfaction is directly influenced by the amount of time required to find the information sought: the less time spent searching, the greater the satisfaction. However, various studies by Hersh and colleagues (e.g. Hersh et al., 1994; Hersh and Molnar, 1995; Hersh and Hickam, 1995; Hersh et al., 1996; Hersh et al., 2000) did not establish any significant relationship between the time needed to complete a search and user satisfaction with the retrieval system.

4.3 H_3 — User effort influences user satisfaction

It is expected that the amount of effort users exert to complete the task influences their satisfaction with a given list of results returned by some search engine: as the amount of effort expended increases, user satisfaction correspondingly decreases. Lancaster (1981: p.113) considered the amount of time the user spent conducting a search as a measure of effort and he also considered the amount of effort expended during the search as a measure of user satisfaction. Kokubu et al. (2005) reported, in a question answering system, an inverse correlation between user satisfaction and the rank position where the answer was located, as users examined more documents by going down the rank to locate relevant information, the less satisfied they were.

4.4 H_4 — User characteristic influences user satisfaction

It was expected that users with previous searching experience were likely to be more satisfied with results than less familiar users. This hypothesis was substantiated by Garoufallou et al. (2007) who found that expert users who have web-search experience are usually significantly more satisfied with the results than novice users as evidenced by the task of utilizing digital heritage maps using the VeriaGrid system. In their study, expert users (this study consisted of

10 participants whose expertise was determined prior to the actual experiment by giving them a list of exercises to be accomplished; based on the proficiency demonstrated when executing these exercises, users were categorized as either experts or novices) took less time to complete the tasks and made fewer errors, thus determining their satisfaction with the results. However, Bruce (1998) did not find any significant effect of previous information search training, frequency of internet usage and reported level of satisfaction of 37 Australian academics. Bruce found these Australian academics to generally have a high expectation of success when they were engaged in information seeking on the Internet, and were satisfied with the process regardless of how frequently they used the Internet or whether or not they had received formal training.

Other studies also noted that users' experience with the IR system under evaluation influenced their satisfaction, particularly if a new technology was presented to them. For example, in their study of CD-ROM database searchers, Steffey and Meyer (1989) concluded that determinants of satisfaction were not associated with actual search success. They ascribed this to potentially confounding variables including experience with computers and fascination with the new database search technology. They noted that "*patrons were so pleased with the electronic periodical indexes, that it did not matter how satisfied they were with the number of citations they had retrieved, or with the value of those citations*." (Steffey and Meyer, 1989: p.43). In their comparison of CD-ROM database searching by professional librarians and university faculty and graduate students, Lancaster et al. (1994) concluded that student and faculty searchers only found about one-third of crucial items. They found that CD-ROM searchers were usually satisfied with less than optimal search results. The authors added: "*Many* [users] *express satisfaction even when they achieve very poor results*." (Lancaster et al., 1994: p.382). They suggest that users are overly impressed with new electronic retrieval technologies, and this accounts for inflated levels of satisfaction with actual search results.

5 Experiment

The experiment reported in this paper was conducted to comprehensively and empirically examine factors (e.g. system effectiveness, user effectiveness, user effort, and user characteristics) influencing user satisfaction. The experiment was motivated by a need to examine the appropriateness of using objective measures such as, system effectiveness to adequately capture the essence of user satisfaction.

5.1 Experiment Methodology

In this experiment, users were presented with systems of varying effectiveness in a controlled way. An experimental test bed (Query Performance Analyzer)[4] was utilized, which allowed access, through a single consistent interface, to three well known experimental retrieval systems (InQuery, Lemur and Terrier). The QPA is a web-based application which uses as a test corpus the TREC-8 document collection. Fifty-six topics were randomly selected from TREC topics; Title and Description fields from each topic were used as queries in the three systems. Each system gave

[4] Query Performance Analyzer http://www.info.uta.fi/julkaisut/pdf/qparn1.pdf [site accessed: 22/02/08].

different results and, using the TREC relevance judgments, we were able to compute Average Precision (AvP)[5] for each system on each topic. For each topic, two systems out of three were selected: the one with the highest AvP score and the one with the lowest AvP score. The system returning the highest number of relevant documents was categorized as the "superior system"; the system returning the fewest, the "inferior system". Therefore, for each topic there were two systems: a superior and an inferior. The difference between the two on each topic provided a measure of system effectiveness against which to compare user satisfaction. Therefore, there was a variation in system effectiveness for each of the 56 topics. Such variation is in itself useful in investigating how user satisfaction is influenced by varying the different levels of retrieval effectiveness. From this design one could investigate whether users were more satisfied with the results retrieved by the superior system as compared to the inferior system. For a more detailed description in the selection of the systems, the reader is advised to refer to Al-Maskari et al.(2008).

5.2 Users

Fifty-six participants were recruited for this experiment; all participants were between the ages of 19 and 40 years old with a median age of 26. Participants were required to engage in a search task that required saving as many relevant documents as possible for a set of 56 TREC topics. Every eight users completed searches on the same set of topics: four in the inferior system and four in the superior system, but the effectiveness of the underlying system was unknown to them. Users were given 7 minutes for each topic, and a Latin-Square arrangement was used to distribute the order of the topics amongst users (to reduce the effects of topic order on results). Users were presented with the description and narrative fields of TREC topics as information needs to be satisfied. They were free to issue multiple queries for each topic within the 7 minutes. The narrative field served as guidance on assessing document relevance using a ternary relevance scheme: highly relevant[6], partially relevant[7] or not relevant. Upon completing a search for each task, users rated their satisfaction with the results in a 4-point scale: very satisfied, partially satisfied, fairly dissatisfied and very dissatisfied.

5.3 Variables Tested in this experiment

Table 1 lists the variables tested in this experiment and their descriptions. Three measures are used to assess experimental results relating to user effectiveness, namely: (i) the number of all relevant documents obtained by the users (including those that both match and do not match with the TREC relevance criteria), (ii) the number of relevant documents identified by the user matching TREC relevance and (iii) the time taken by the user to locate the first relevant document which matches with TREC relevance assessment. The reason for considering documents that match and do not match the TREC relevance assessment is because users submitted their own queries and there were documents retrieved which did not have TREC relevance judgments, thus it was necessary to adopt the user's judgments of relevance for those documents lacking these TREC relevance scores.

[5] Average precision obtained after each relevant document is retrieved.

[6] The document directly addresses the core issue of the topic.

[7] The document only points to the topic: it does a not discus the themes of the topic thoroughly.

Variable Name	Description
UserDocs	Represents the number of all relevant documents obtained.
TRECDocs	Represents the number of relevant documents obtained which are also consistent with TREC relevance assessments.
TimeFRD	Represents the time taken by the users to locate the first relevant document that match with TREC relevance assessment for a given search task.
NumQrs	Indicates the number of queries submitted to complete the task.
RankFRD	Represents the rank position in the results list accessed to obtain the first relevant document that match with TREC relevance assessment.
UsrSat	A score representing the user's level of satisfaction with the Results.

Table 1: *Variable Names and Their Descriptions.*

User effort was quantified using two criteria, namely: (i) the number of queries submitted, and (ii) the lowest rank position in the result list accessed by the user to locate the first relevant document. It is to be noted that, for the purposes of the experiment discussed in this paper, any form of query modification made by users to their initial queries (substituting a synonym for some item in the query term, correcting spelling errors, hyphenating, etc.) was classified as a new query for the topic being explored.

5.4 Results

This section was divided into four sub-sections: i) the effect of system effectiveness on user satisfaction; ii) the effect of user effectiveness on user satisfaction; iii) the effect of user effort on user satisfaction; and v) the effect of user characteristics on user satisfaction.

6.5 The Effect of System Effectiveness on User Satisfaction H_1

Table 2 presents the effectiveness of both the inferior and superior system (as measured by MAP[8]); this table also presents the scores of user effectiveness, user effort and user satisfaction while conducting search operations in these systems. As illustrated in Table 2, in general, users were significantly more satisfied with the results of the superior system than those of the inferior system. These results suggested that when users were given different systems with a substantial difference in their effectiveness, it was likely that users were more satisfied when executing search operations on a more effective system than a less effective system. A significant (spearman) correlation was found between system effectiveness and user satisfaction ($r = 0.269^{**}$). However, this correlation was not strong which perhaps indicated that user satisfaction was not easy to predict when correlated with system effectiveness.

[8] The mean of the precision scores obtained after each relevant document is retrieved.

Measures	Inferior system	Superior System	Significance (t-test)
MAP	0.05	0.20	0.00**
UserDocs	3.67	5.45	0.00**
TRECDocs	2.00	3.37	0.00**
TimeFRD	2.39	1.79	0.02*
NumQrs	4.09	3.27	0.00**
RankFRD	8.37	3.11	0.01**
UsrSat	2.04	2.34	0.00**

** $p<0.01$; * $p<0.05$

Table 2: *Users' effectiveness and their satisfaction of the inferior and superior system.*

5.6 The Effect of User Effectiveness on User Satisfaction H_2

This section discusses H_2 which posited that user effectiveness directly impacted upon user satisfaction, i.e. the higher the number of relevant documents obtained; the higher the user satisfaction. A spearman correlation indicated a significant relationship between *UsrSat* and the following criteria: (i) *UserDocs* score ($r = 0.604^{**}$), (ii) *TRECDocs* score ($r = 0.550^{**}$), and (iii) *TimeFRD* score ($r = -0.293^{**}$).

Figures 2, 3 and 4 depict user satisfaction with the results in relation to *UserDocs*, *TRECDocs* and *TimeFRD* scores. The experimental results visualized in these figures show that user satisfaction fluctuated in direct relation to user effectiveness which supports H_2.

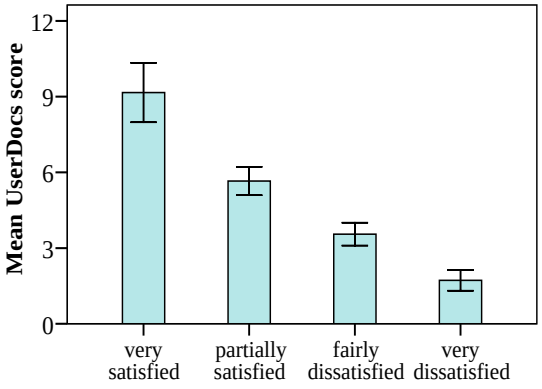

Figure 2: *Users' satisfaction vis-à-vis UserDocs score.*

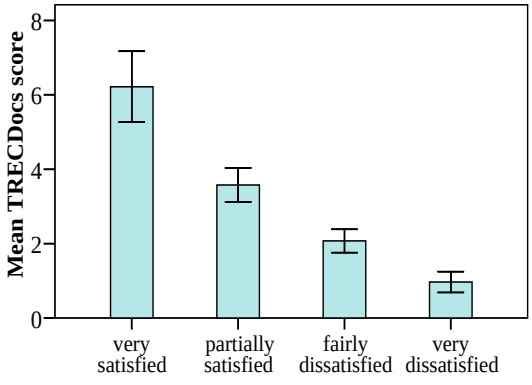

Figure 3: *Users' satisfaction vis-à-vis TRECDocs score.*

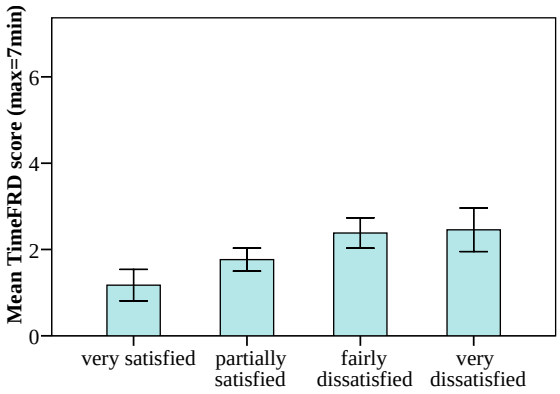

Figure 4: *User satisfaction in reference to TimeFRD score.*

5.7 The Effect of User Effort on User Satisfaction H_3

This section further explores the relationship between user effort (the *NumQrs, RankFRD*) and user satisfaction. An inverse Spearman's correlation was observed between the *NumQrs* and *UsrSat* score (r = -0.29**) and between the *RankFRD* and *UsrSat* scores (r = -0.19**). These inverse correlations indicate that as users exert more effort, the less satisfied they become. Figures 5 and 6 show participant satisfaction with results returned by search engines in response to *NumQrs* and *RankFRD*. As the *NumQrs* score increased and the searcher was obliged to access documents at lower rank positions *(RankFRD)*, there was a corresponding decrease in user satisfaction. In summary, results from this section supported the assumption made in H_3 concerning the relationship between user effort and user satisfaction.

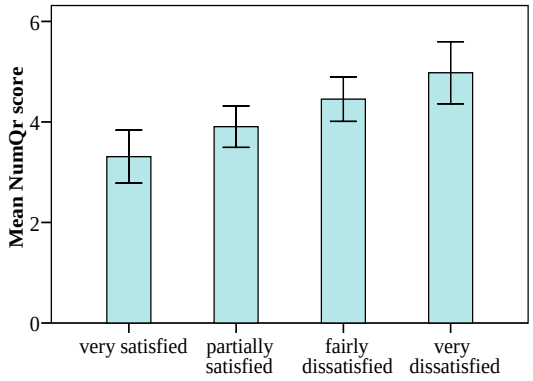

Figure 5: *Users' satisfaction in relation to NumQrs score.*

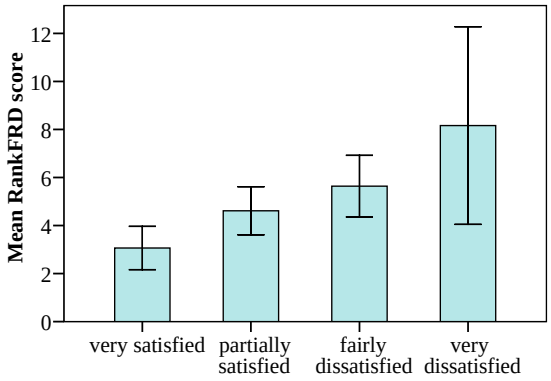

Figure 6: *Users' satisfaction in relation RankFRD score.*

5.8 The Effect of User Characteristics on User Satisfaction H_4

This section examined the relationship between user characteristics and user satisfaction, an example of such being the effect of search topic familiarity on user satisfaction. It was expected that 'familiar' users would be more satisfied with the results than less familiar users (Figure 7). Nonetheless, there was no relationship between

familiarity and satisfaction. This lack of correlation between user familiarity and user satisfaction was not surprising because a user's familiarity with a given search topic had no impact upon his/her effectiveness. It was found that there was no significant difference in *UserDocs*, *TRECDocs* and *TimeFRD* scores between familiar and unfamiliar users.

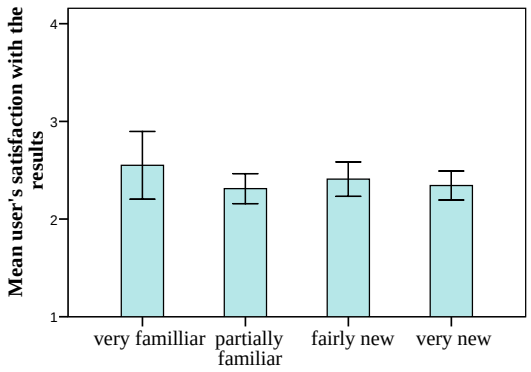

Figure 7: *Users' search topic familiarity vis-à-vis satisfaction with search results.*

Another user characteristic under review in this section was online search experience. The results in Figure 8 indicated that experienced and inexperienced users were equally satisfied with the results. Therefore, results from this section did not support H_4. It was assumed that familiar users were more effective than unfamiliar users and therefore would be more satisfied with the results of their search. On the contrary, it was found that familiar users were found to be as effective as unfamiliar users and therefore as satisfied.

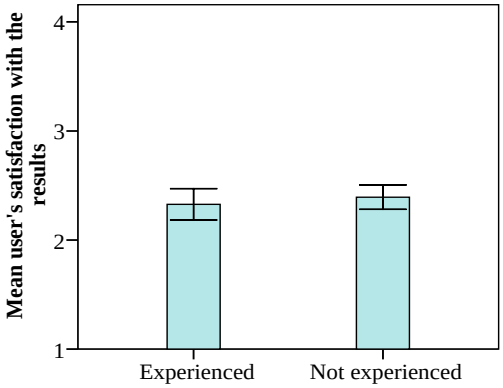

Figure 8: *Users' satisfaction in relation to online search experience.*

6 Discussion

In the beginning of this paper, four factors were listed that were expected to influence user satisfaction: system effectiveness, user effectiveness, user effort, and user characteristics. In the following is a summary of the influence of these factors on user satisfaction.

6.1 System Effectiveness

It was speculated in H_1 that system effectiveness influences user satisfaction; for example an improvement in system effectiveness results in an increase in user satisfaction. Results from the experiment (as detailed in section 6) verified that system effectiveness had a positive correlation with user satisfaction; however, this correlation was not strong (r=0.27**). The low correlation between system effectiveness and user satisfaction may be due to an evaluation discrepancy among users, i.e. some users were more generous in awarding high grades for satisfaction than others. The finding that system effectiveness influenced user satisfaction was consistent with previous studies (e.g. Gluck, 1996; Johnson et al., 2003; Thomas and Hawking, 2006; Huffman and Hochster, 2007) that observed a correlation between the relevance of results and user satisfaction.

6.2 User Effectiveness

It was assumed in H_2 that user effectiveness influenced their satisfaction. Results from the experiment established a relationship between user effectiveness (number of relevant documents obtained) and their satisfaction (r = 0.55**). It was also found that user satisfaction was directly influenced by the amount of time required to find the information sought: the less time spent searching, the greater the satisfaction. Previous studies (Su, 2003; Law et al., 2006) also concluded that user satisfaction was directly influenced by the amount of time required to find the information sought: the less time spent searching, the greater the satisfaction.

6.3 User Effort

H_3 proposed that user effort had a degree of influence on user satisfaction. This hypothesis was supported by the results of the experiment; it was observed that user effort – as measured by the number of queries submitted to obtain relevant documents and the rank position in the results list accessed to obtain the relevant documents – was inversely correlated with user satisfaction. As users exert an increasingly greater effort to complete a given search task, it was very likely that their satisfaction decreased. Kokubu et al. (2005) also reported, in a question answering system, an inverse correlation between user satisfaction and the rank position where the answer was located, as users examined more documents by going down the rank to locate relevant information, the less satisfied they were.

6.4 User Characteristics

The premise of H_4 was that user characteristics influenced user satisfaction. For the purposes of this investigation, user characteristics were defined and measured according to: i) familiarity with the search topic, and ii) search experience. Results from the experiment reported in this paper did not reveal any correlation between user characteristics and user satisfaction. It was expected that familiar users with the search topics would be more effective (and therefore more satisfied) than unfamiliar users. However, the experimental findings in this case showed that unfamiliar users were as satisfied as familiar users.

However, previous studies (e.g. Garoufallou et al., 2007) substantiated a relationship between user search experience and user satisfaction; they found that users who had web-search experience were usually significantly

more satisfied with the results than novice users. In their study of CD-ROM database searchers, Steffey and Meyer (1989) and Lancaster et al. (1994) found that experience with the IR system under evaluation influence users rating of their satisfaction because it lead to fascination with the new search technology. Although from this experiment we did not observe any effect of user characteristics on user satisfaction, we believe that these findings are not conclusive and further investigation is need on this matter. Given that previous studies indicated that certain user characteristics (i.e. experience and knowledge of the system under evaluation) have an impact on user satisfaction. For example, we assume that users may express high satisfaction and be impressed with certain features offered by the IR system if they are exposed to them for the first time and thus satisfaction level might change over time.

The individual nature of user satisfaction makes its measurement a non-trivial proposition since it is based largely on the knowledge constraints for a given user, i.e. what that user knows and does not know (Lancaster, 1981: p.106). For example, a user expresses satisfaction with the results based on what is known to him. If all the retrieved items are relevant to a user, then this user will indicate high satisfaction; however, the user might be less satisfied with the results if he knew that many relevant items are missed from the results, in particular is the missed items are more relevant than the retrieved items. Hufnagel (1990) also questioned the validity of referring to user satisfaction ratings when evaluating system effectiveness. He argued that user satisfaction ratings may be a reflection of performance outcomes (i.e. the success or failure) for individual search tasks. However, results from this paper support using user satisfaction to infer system effectiveness, this is based on the significant correlation observed between system effectiveness and user satisfaction. Furthermore, it was illustrated that while users searching in superior and inferior systems they are significantly more satisfied with the system with the superior effectiveness as compared to their satisfaction with the inferior effectiveness.

7. Conclusion

It was stated in the beginning of this paper that there is a debate in the literature concerning the applicability of user satisfaction as a measure of system effectiveness. We contend that the results from the experiment reported in this paper provides a clearer picture on the relationship between user satisfaction and system effectiveness. It was illustrated that users were significantly more satisfied with a system having higher effectiveness compared to a less effective system.

Previous studies that examined user satisfaction (e.g. Sandore, 1990; Gluck, 1996; Turpin and Hersh, 2001; Johnson et al., 2003; Huffman and Hochster, 2007) only looked at one factor (e.g. the relationship between user satisfaction and system effectiveness or the relationship between user satisfaction and user effectiveness) whereas this study was more comprehensive in that it examined all the possible factors that could influence user satisfaction. We believe that one should not consider one factor per experiment, because user satisfaction is easily affected by other factors which can confound the results and affect the conclusion drawn from an experiment.

In conclusion, it has been illustrated that user satisfaction is a subjective variable and it is influenced by several factors (e.g. system effectiveness; user effectiveness, user effort and user characteristics). However, it must be emphasised that when users do not have previous experience in the IR system under evaluation, it is likely that the user group's expectations are low concerning what to expect in terms of the search engine's quality of output

cause to very little variation in the users' subjective ratings of their satisfaction. Therefore, IR system evaluators should consider all factors listed above in measuring user satisfaction. To accurately measure user satisfaction, for instance, it is not adequate to say that system effectiveness do not influence user satisfaction; we also need to consider user experience with the IR system under evaluation. Future work to be carried out includes examining the factors that are found to influence user satisfaction using different IR systems than the one used in this experiment (e.g. different queries and with more modern search systems).

References

Al-Maskari, A., Sanderson, M. & Clough, P. (2008)."The Good and the Bad System: Does the Test Collection Predict Users' Effectiveness". In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp.59-66 Singapore: ACM.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999) Modern Information Retrieval, Boston, MA, Addison Wesley Longman

Belkin, N. J. & Vickery, A. (1985) Interaction in information systems: a review of research from document retrieval to knowledge-based systems, London, British Library.

Bruce, H. (1998) User satisfaction with information seeking on the Internet. *Journal of the American Society for Information Science,* **49** (6)**,** 541-556.

Cooper, W. S. (1968) Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation,* **19** (1)**,** 30–41.

Cooper, W. S. (1973) On selecting a measure of retrieval effectiveness part II. Implementation of the philosophy. *Journal of the American Society for Information Science,* **24** (6)**,** 413-424.

Egan, D. E. (1988) Individual differences in human-computer interaction. IN HELANDER, M. G., LANDAUER, T. K. & PRABHU, P. V. (Eds.) *Handbook of human-computer interaction* Amsterdam, Holland, Elsevier.

Frøkjær, E., Hertzum, M. & Hornbæk, K. (2000)."Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated?" In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp.345 - 352 The Hague, Amsterdam: ACM.

Garoufallou, E., Trohopoulos, I., Siatri, R. & Balatsoukas, P. (2007)."A user-centred approach to the evaluation of digital cultural heritage maps: the VeriaGrid system ". In: *Libraries Without Walls 7,* Aegean Island of Lesvos, Greece:

Gatian, A. W. (1994) Is user satisfaction a valid measure of system effectiveness? *Information and Management,* **26** (3)**,** 119-131.

Gluck, M. (1996) Exploring the relationship between user satisfaction and relevance in information systems. *Information Processing and Management,* **32** (1)**,** 89-104.

Griffiths, J., Johnson, F. & Hartley, R. (2007) User satisfaction as a measure of system performance. *Journal of Librarianship and Science,* **39** (3)**,** 142-152.

Harter, S. P. & Hert, C. A. (1997) Evaluation of information retrieval systems: Approaches, issues, and methods. *Annual Review of Information Science and Technology (ARIST),* **32,** 3-94.

Hersh, W., Buckley, C., Leone, T. J. & Hickam, D. (1994)."OHSUMED: an interactive retrieval evaluation and new large test collection for research". In: *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval* pp.192-201. Dublin, Ireland: Springer-Verlag

Hersh, W. & Hickam, D. H. (1995) An evaluation of interactive Boolean and natural language searching with an online medical textbook. *Journal of the American Society for Information Science (JASIS),* **46** (7)**,** 478-489.

Hersh, W., Pentecost, J. & Hickam, D. (1996) A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science,* **47** (1)**,** 50-56.

Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L. & Olson, D. (2000)."Do batch and user evaluations give the same results?" In: *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval* pp.17-24. Athens, Greece: ACM.

Hildreth, C. R. (2001) Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study. *Information Research.*

Huffman, S. B. & Hochster, M. (2007)."How well does result relevance predict session satisfaction?" In: *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval* pp.567-574. Amsterdam, The Netherlands: ACM.

Hufnagel, E. M. (1990)."User satisfaction-are we really measuring system effectiveness". In: *Proceedings of the 23rd Annual Hawaii International Conference on System Science*, pp.237-446. Hawaii, USA: IEEE Computer Society

Järvelin, K. & Kekäläinen, J. (2000)."IR evaluation methods for retrieving highly relevant documents". In: *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval*, pp.41 - 48 Athens, Greece: ACM.

Johnson, F. C., Griffiths, J. R. & Hartley, R. J. (2003) Task dimensions of user evaluations of information retrieval systems. *Information Research.*

Kokubu, T., Sakai, T., Saito, Y., Tsutsui, H., Manabe, T., Koyama, M. & Fujii, H. (2005)."The Relationship between Answer Ranking and User Satisfaction in a Question Answering System". In: *Proceedings of the 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-5)*, pp.537-544. Tokyo, Japan:

Korfhage, R. R. (1997) Information Storage and Retrieval, New York, John Wiley & Sons.

Lancaster, F. W. (1969) MEDLARS: report on the evaluation of its operating efficiency. *American Documentation,* **20** (2)**,** 119-142.

Lancaster, F. W. (1979) Information retrieval systems: characteristics, testing and evaluation, New York ; Chichester (etc.), John Wiley & Sons.

Lancaster, F. W. (1981) Evaluation within the environment of an operating information service. IN SPÄRCK JONES, K. (Ed.) *Information Retrieval Experiment.* London, Butterworths.

Lancaster, F. W., Cheryl, E., Mary, J. Z., Laura, M. & Yuen-Man, L. (1994) Searching Databases on CD-ROM: Comparison of the Results of End-User Searching with Results from Two Modes of Searching by Skilled Intermediaries. *RQ,* **33** (3)**,** 370-386.

Law, E. L.-C., Klobučar, T. & Pipan, M. (2006) User Effect in Evaluating Personalized Information Retrieval Systems. IN NEJDL, W. & TOCHTERMANN, K. (Eds.) *LNCS 4227.* Springer Berlin / Heidelberg.

Marchionini, G. (1995) Information Seeking in Electronic Environments Cambridge University Press.

Nielsen, J. (1993) Usability Engineering, San Francisco, Morgan Kaufmann.

Sandore, B. (1990) Online Searching: What Measure Satisfaction? *Library and Information Science Research,* **12** (1)**,** 33-54.

Saracevic, T. & Kantor, P. (1988) A study of information seeking and retrieving. II. Users, questions, and effectiveness. *Journal of the American Society for Information Science,* **39** (3)**,** 177-196.

Soergel, D. (1976) Is user satisfaction a hobgoblin?' *Journal of the American Society for Information Science,* **27** (4)**,** 256-259.

Spärck Jones, K. (1981) Information Retrieval Experiment, London, Butterworths.

Steffey, R. J. & Meyer, N. (1989) Evaluating User Success and Satisfaction with CD-ROM. *Laserdisk Professional,* **2** (5)**,** 35-45.

Su, L. T. (1991). An investigation to find appropriate measures for evaluating interactive information retrieval. PhD, Rutgers University

Su, L. T. (1992) Evaluation measures for interactive information retrieval. *Information Processing and Management,* **28** (4)**,** 503-516.

Su, L. T. (1998) Value of search results as a whole as the best single measure of information retrieval performance. *Information Processing and Management,* **34** (5)**,** 557-579.

Su, L. T. (2003) A comprehensive and systematic model of user evaluation of Web search engines: II. An evaluation by undergraduates. *Journal of the American Society for Information Science and Technology,* **54** (13)**,** 1193-1223.

Tessier, J. A., Crouch, W. W. & Atherton, P. (1977) New Measures of User Satisfaction With Computer Based Literature Searches. *Special Libraries,* **68** (11)**,** 383-389.

Thomas, P. & Hawking, D. (2006)."Evaluation by Comparing Result Sets in Context". In: *Proceedings of the conference on information and knowledge management (CIKM)*, pp.94-101 Arlington, Virginia, USA: ACM.

Turpin, A. H. & Hersh, W. (2001)."Why batch and user evaluations do not give the same results". In: *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval*, pp.225 - 231 New Orleans, Louisiana, United States ACM