Information Research, Vol. 8 No. 4, July 2003

Task dimensions of user evaluations of information retrieval systems

F.C.Johnson, J.R. Griffiths and R.J. Hartley
Department of Information and Communications
Manchester Metropolitan University
Manchester, UK

Abstract

This paper reports on the evaluation of three search engines using a variety of user-centred evaluation measures grouped into four criteria of retrieval system performance. This exploratory study of users' evaluations of search engines took as its premise that user system success indicators will derive from the retrieval task the system supports (in its objective to facilitate search). This resulted in the definition of user evaluation as a multidimensional construct which provides a framework to link evaluations to system features in defined user contexts. Our findings indicate that users' evaluations across the engines will vary, and the dimensional approach to evaluation suggests the possible impact of system features. Further analysis suggests a moderating effect on the strength of the evaluation by a characterization of the user and/or query context. The development of this approach to user evaluation may contribute towards a better understanding of system feature and contextual impact on user evaluations of retrieval systems.

Introduction

This paper presents the development of a framework for the evaluation of Information Retrieval (IR) systems, focusing on Internet search engines from a user perspective. The traditional measures for evaluation based on the relevancy of the retrieved output may only be a partial match of users' objectives and of the systems' objectives. The user's judgement of the success of the system may be influenced by factors other than the recall and precision of the output. Such factors are likely to be related to the degree to which the system meets *its* objective to facilitate and maximize the users' search. Usability measures are typically used for the evaluation of an interactive retrieval system and allow some investigation of the impact of system features on users' search behaviour and, in turn, system performance. An alternative approach is to substitute (action based) usability measures for users' assessment of the system on selected variables. Towards this end we posit that users' evaluation of a retrieval system is a multidimensional construct based on the user information searching process which the system seeks to support. In this paper we investigate the development of this criterion on which users might judge the success of a retrieval system. A small scale evaluation of three search engines was carried out using groups of possible success indicators and the interrelations between these variables were examined to suggest those which appear to be determinants of users' judgment of system success (with information retrieval). This initial investigation provides a basis on which we can speculate the value of dimensional user evaluations of system success defined in terms of system suitability for the user task. It is proposed, for further research, that this user-centred approach can provide a framework in which user evaluations on the dimensions can relate to relevant system and situational impacts. If users' evaluations, determined by the search process, are dependent on system features then we can expect evaluation to vary across systems. Further, if a moderating effect of the user query context can be observed then we may be closer to understanding variations in users' evaluations of the same system.

Search engine evaluation and development

Harter and Hert (1997) in their comprehensive review of evaluation of IR systems define evaluation partly as a process by which system effectiveness is assessed in terms of the degree to which its goals and objectives are accomplished. Modern interactive retrieval engines aim to support the user in the provision of features and functionality which help maximize their search for the retrieval of relevant and pertinent results. The user and their search as an integral component of the system under investigation has been the focus of ongoing developments in evaluation methodologies for interactive retrieval systems (for example, Robertson and Hancock-Beaulieu, 1992 and Harman, 2000. As Dunlop (2000: 1270), reflecting on MIRA, a working group to advance novel evaluation methods for IR applications, states:

> the challenge for interactive evaluation in IR is to connect the two types of evaluation: engine performance, and suitability for end-users.

Traditional recall and precision measures of engine performance are based on the concept of relevance, that for a given query there is a set of documents which match the subject matter of that query. Relative recall and precision measures have been used for the evaluation of search engine performance but comparison across these studies is difficult. Different scales for the relevance judgements have been used (Chu and Rosenthal, 1996; Ding and Marchionini, 1996; Leighton and Srivastava, 1999) and/or the results are drawn from different query sets (Gauch and Wang, 1996; Tomaiuolo and Packer, 1996; Back, 2000). It was, however, the long standing criticism of the very basis of these performance measures, binary relevance judgments, which was highlighted as a major criticism of the evaluations of web search engines participating in the Web Special Interest track of the large scale testing environment of the Text Retrieval Conferences (TREC-8) (Hawking *et al.*, 1999; Hawking *et al.*, 2001).

Sherman's report of the Infonortics 5th search engine meeting (2000) states that the relevancy of the system output is a poor match of these systems' objectives which he suggests include getting information from users, providing browsing categories, promoting popular sites and speed of results. Those, it seems, which assist a user to express a query, navigate through the collection, or quickly get to the requested information. This trend in the development of search assistance features supporting casual users has been consistently noted (Sullivan, 2000; Feldman, 1998, 1999; Wiggens and Matthews, 1998). Search and retrieval features are touted as maximising user capabilities in manipulating the information searching process. Statistical probablistic retrieval systems allow natural language or, more precisely, unstructured queries which may support the user's task of query formulation; concept processing of a search statement may determine the probable intent of a search; relevance feedback can assist users in modifying a query; and, use of on and off the page indicators to rank the retrieved items may assist users in judging the hit list, as do visualisation techniques, for example, which may be used to provide a view of a subset of the collection.

The increasingly interactive nature of system design motivates approaches to evaluation which accommodate the user and the process of interaction. Front-end features will affect users' interactions which in turn will partly determine the effective performance of the back-end system. The requirements which can be derived from this interaction of system features, user and query are articulated well in Belkin *et al.*, as highlighted in Harter and Hert, 1997,: 26

> if we are going to be serious about evaluating effectiveness of interactive IR, we [sic. need to] develop measures based upon the search process itself and upon the task which has lead the searchers to engage in the IR situation.

The rise of interactivity was acknowledged from TREC-3 onwards with the introduction of a special interactive track (Beaulieu *et al.*, 1996) and its goal to investigate the process as well as the outcome in interactive searching (Hersh and Over, 2001) Various usability measures of the user-system performance, such as number of tasks completed, number of query terms entered, number of commands used, number of cycles or query reformulations, number of errors and time taken, were thus derived by consideration of the users' actions or behaviour in carrying out a search task. As such, usability measures may provide indicators of the impact specific features of a retrieval system have on searcher behaviour. Voorhees and Garofolo, (2000) itemise studies which have investigated features such as the effects of visualization techniques, and different styles of interaction, while several studies have focused on the system feature of relevance feedback in interactive IR. Belkin *et al.* (2001) looked at query reformulation, and White *et al.* (2001) compared implicit and explicit feedback. In the context of TREC-8 interactive track, Fowkes and Beaulieu (2000) examined searching behaviour, related to the query formulation and reformulation stages of an interactive search process, with a relevance feedback system. Of particular note in this study was the moderating effect of the user query context, where different query expansion techniques were found to be suited to the degree of query complexity.

Others substitute usability measures for satisfaction measures and ask users about their satisfaction with or judgment of the general performance of the system and possibly its specific features (or more widely, the interface). Su (1992, 1998) identified twenty user measures of system success which were grouped into the evaluation dimensions of relevance, efficiency, utility and user satisfaction. These were correlated to the users' rating for overall system success to determine 'value of search results as a whole' to be the best single user measure of success. The pursuit to find the best determinant of users' system success rating is a worthy one which would reduce considerable cost and effort in evaluation. Yet its application would mask the potentially complex judgment resulting from the many factors with which a user engages in interacting with the system and on which the user may draw in assessing the system. The crux of the issue being that while a user evaluation (expressed in a single construct) is more easily obtained, it is at the expense of knowing why. From the system developers' perspective it may be of value to have greater insight into user evaluations and the subtle balance of the interrelations between the various indicators of a user judgment of system success. A single system, for example, may be rated as providing excellent results but requiring considerable user time and effort, thus high on effectiveness but low in efficiency or indeed visa versa. Hildreth (2001), for example, found that 'perceived ease of use' of an OPAC was related to 'user satisfaction with the results'. Our concern with what might be the determinants of a user judgment such as 'ease of use' leads to consideration of a more detailed system evaluation. For example, a system's layout and representation of the retrieved items may contribute to a user perception of ease of use helping the user to quickly and easily make relevancy judgments. It would seem reasonable to speculate that a system could score high in the user's judgment of this aspect but, in fact, score low on user satisfaction with the search results.

Dimensions and determinants of user evaluation.

A system evaluation based on various success indicators requires some basis on which the interrelations between these factors can be sought and understood. This requirement is perhaps demonstrated by the fact that previous studies have been unable to establish a consistent relation between user satisfaction and the recall and precision of the search results (Sandore, 1990; Su, 1992; Gluck, 1996; and Saracevic and Kantor, 1988). The reason for this may be that it is an erroneous assumption that there ought to be a relation between user and system measures of performance. It would certainly be dependent on the individual and their situation. The framework we seek to develop is thus one which relates and groups various user indicators of success into dimensions of the users' overall judgment in order that meaningful relations can be sought in the evaluation. The remainder of this paper describes our preliminary investigation into the defining function of the retrieval task dimensions on user evaluation indicators and measures. Ideally, user evaluation based on what the user is doing and the system is supporting will link to system features in well defined search contexts. A further requirement for our feasibility study is thus to identify possible user/query contexts which may moderate users' judgments of the system. Users with different demands on the system are likely to vary accordingly in their assessment based on the system and its features. Succinctly put, our feasibility study set out to identify the possible dimensions and indicators of user evaluations in a framework in which evaluation is dependent on system features and moderated in a user/query context.

The IR task process

Several models of information searching have been suggested which may be used to identify the user success factors in system evaluation. Information retrieval, for example, may be viewed as an exploratory process, a view which led Brajnik (1999) to derive evaluation statements for criteria such as system flexibility. The approach we explored focused on the more mechanical process models to demarcate the tangible (system dependent) information retrieval activities. Such models of the basics of information searching are standard, for example in Salton (1989) as highlighted in Baeza-Yates (1999: 262), and from which the following interacting steps are defined:

- formulation and submission of a query,
- examination of the results, *with a*
- possible feedback loop to re-formulate the query, *and*
- integration of search results and evaluation of the whole search.

Each step provides some statement of user-requirement, what the goal-directed user is trying to do with the system. Each step in the process thus represents a dimension which defines the variables and measures on which a user may evaluate a system's success in supporting information retrieval.

Table 1 groups these indicators of system success under each of the dimensions, Query formulation, Query reformulation, Examination of result, and Evaluation of search results and search as a whole. The majority of which came from existing (and generally accepted) measures of retrieval system performance (for example as listed in Su (1992) and which form the broad criteria of effectiveness, utility and efficiency. The main effect of using the lower level task dimensions is the decomposition of interaction by the three task process steps.

Effectiveness	Utility	Efficiency
Evaluation of results Satisfaction with precision Satisfaction with ranking	**Evaluation of search and results** Value of search results Satisfaction with results Resolution of the problem Rate value of participation Quality of results	**Evaluation of search as a whole** Search session time Response time
Interaction		
Query formulation Satisfaction with query input	**Query reformulation** Satisfaction with query modification Satisfaction with query visualisation	**Examine results** Satisfaction with visualisation of item Satisfaction with manipulation of output

Table 1: Indicators of user judgment of system success grouped by task dimensions

Measures of search results

Users' judgment of the success of the search results and the search as a whole may be based on the criterion of effectiveness manifest in the system output, the retrieved set. Traditional measures of retrieval effectiveness are based on the notion of relevance, and evaluation of the system will be partially dependent on the ability of the system to meet its basic function to retrieve only relevant documents. We used Su's user measures of user satisfaction with precision, and ranking. Further indicators of a user success judgment of the search may be based on factors, other than relevance. Measures of utility focus on the actual usefulness or value of the retrieved items to the individual information seeker (Saracevic *et al.*, 1988). Cleverdon (1991) argued (with Cooper, 1973, who put forward a straight utility-theory single measure) that retrieval effectiveness measures of recall and precision should be used in combination with (and possibly related to) these more user-oriented measures which are based on factors as subjective satisfaction statements, search costs, and time spent. Indeed, various factors may bear on users' judgements of overall satisfaction with the value of the search results. For example, users may be influenced by the extent to which information quality can be assumed based on the source; the extent to which the information is accurate or correct; and, the extent to which the information is at the right level to meet user need. Yet, Saracevic (and Su, 1998: 558) report that standard utility measures do not exist. Saracevic used the following evaluative statements:

- How much time spent reviewing abstracts;
- Assign a cost value to usefulness of results;
- What contribution this information made to resolution of problem that motivated your question;
- Overall how satisfied with results.

Su found that a measure of utility **value of the search results as a whole** correlated most strongly with users' overall judgement of system success and thus proposed it to be a best single measure of a system. We used as a measure of utility *Satisfaction with results* and *resolution of the problem* (derived from Saracevic) and *value of search results as a whole*, *value of participation* and *quality of results* (from Su).

Measures of user-system interaction

In the process of getting the search results the user is involved in the steps of query formulation, re-formulation and examination of the results. These comprise broad categories of user interaction with the system, and the specific actions subsumed relate to how the user interacts and manipulates or commands the system to retrieve the required information. In the absence of usability measures, interaction will be largely determined by satisfaction measures alone. As Belkin and Vickery (1985: 194) point out satisfaction is a concept intended to capture an overall judgement based on user reaction to the system thus extending the range of factors relevant to the evaluation. Our approach which has interaction as a broad category of the three interactive task dimensions breaks down the concept to measures based on these relevant factors, grouped in Table 1 and explained as follows. **Query formulation**, on consulting an information retrieval system, user reaction to the system may be influenced by the perceived ease of expressing the query. The user may be influenced in this judgement by, for example, the availability of different search methods, such as natural language searching or power search to specify a search topic. A measure of *user satisfaction with query input* thus is defined in terms of the perceived ease in the expression or specification of the query. **Search reformulation**, users may be influenced by any assistance or feedback received for formulating or modifying the query, such as the system suggesting query terms from a thesaurus or offering 'more like this' feedback options. Another form of feedback lies with the query visualization, the assistance provided in understanding the impact of a query. An example is the use of folders to categorise search results which may suggest to the user different perspectives of the topic which is useful in refining the search. The measures of *user satisfaction with query modification* and *satisfaction with query visualization* are thus defined in terms of system suggesting search terms or facilitating query by example, and in terms of understanding the impact of the query respectively. **Examining the results**, on receiving results the user will be involved in some process of interpreting the results in the given frame of the information need and would want to see quickly and easily an item's topic or meaning and why it was retrieved. Summary representation features for visualising the 'aboutness'of an item might support a user in this task, e.g., in highlighting query terms, showing category labels and a clear and organised layout. Thus we defined the *measure of satisfaction with visualisation of item representation* and *manipulation of the output* (e.g., summary display features (category labels), sort by).

Measures of search efficiency

The final stage in the retrieval task is evaluation of the search as a whole and relates to the criterion of efficiency. Boyce *et al.* (1994: 241) highlight the difference between effectiveness and efficiency thus:

> an effectiveness measure is one which measures the general ability of a system to achieve its goals. It is thus user oriented. An efficiency measure considers units of goods or services provided per unit of resources provided.

Dong and Su (1997: 79) state that response time is becoming a very important issue for many users. If users want to retrieve information as quickly as possible, this may in part equate to the efficiency of the system and the judgement of which will affect user evaluation of the system success as a whole. Thus whilst efficiency seems hard to define, these studies and others (such as, Stobart and Kerridge, 1996; Nahl, 1998) seem corroborate on the importance of speed of response as an indicator of efficiency.

Multidimensional framework for search engine evaluation

In our feasibility study these indicators of satisfaction were used to elicit a user judgment of the success of three search engines. Our aim was not to obtain an evaluation of these engines as such but rather to confirm or refute our premise that a dimensional user evaluation can result in meaningful relations sought in the user judgments made across and within the systems. This can be stated in three propositions which were explored to varying degrees afforded by the scope of the study.

Proposition One states that the user judgment of system success is a response to how well the system has supported the retrieval task and, as such, is a multidimensional construct. The nature of user evaluations was explored as follows:

- User success ratings assigned on the four criteria were correlated with an overall success judgment to find which, if any, appears be the most important factor in defining user judgment of the system.
- Users ratings on the measures were correlated with the overall success ratings for each associated criterion to find which, if any, contributed most strongly to the user's overall rating of a criterion.
- User derived reasons for attributing satisfaction ratings, overall and on each criterion, were collected using

open-ended questions and analysed to suggest, or otherwise, our measures as those which users themselves base an evaluation of system success.

Proposition Two states that user ratings will vary across the systems reflecting the support of system features to the retrieval task. In the scope of this feasibility study we do not claim to test this, but evidence of it was sought in the finding that user evaluations varied across the engines.

- User ratings on each of the measures were compared across the search engines to find which engine, if any, received notably higher/lower ratings. Some speculation was made as to the possible impact of system features.

Proposition Three states that user evaluations of the systems will be moderated by the context of the users' information query making different demands on the system. Again this was not tested but evidence was sought of contexts leading to the systems receiving different evaluations.

- Four task identifiers were analysed against the overall user ratings and the four evaluation criteria across all four search engines to ascertain if a moderating effect of context was obtained.

Implementation

Twenty-three participants were recruited from second year students of the Department of Information and Communications, MMU. A short briefing was given a few days prior to their search session to explain the project and to present the Information Need Characteristic Questionnaire which was to be completed before the search session. No restrictions were placed on the type of information sought or the purpose for which it was intended, but the questionnaire did capture a characterisation of the search context in terms of following parameters used in Saracevic (1988): a) problem definition (*on a scale from 1-5, would you describe your problem as weakly defined or clearly defined?*); b) intent (*on a scale from 1-5, would you say that your use of this information will be open to many avenues, or for a specifically defined purpose*); c) amount of prior knowledge (*on a scale from 1-5, how would you rank the amount of knowledge you possess in relation to the problem which motivated the request?*); and d) expectation (*on a scale from 1-5 how would you rank the probability that information about the problem which motivated this research question will be found in the literature?*). Participants were then asked to conduct the search using as many reformulations as required and to search for as long as they would under normal conditions. Each was required to search the three engines, Excite, NorthernLight, and HotBot the order of which was varied to remove learning curve effect. These engines were selected on the basis that each had at least one discernable feature so that each search would present a unique search experience and the ability to distinguish the engines.

Following each search participants were required to respond on a likert type scale to questions relating to each of the user satisfaction variables indicated in **Table 1** as defining each of the evaluation criterion. In addition users were asked to provide an overall success rating of the engine with respect to each criterion, i.e., effectiveness, efficiency, utility and interaction. This was to allow the testing of Proposition one as stated above towards the identification of the measures which when validated could form the basis to comprise a judgment for each criterion.

To measure system effectiveness users were asked to rate on a three point scale the degree of relevance of each item retrieved, leaving it open as to how many individual items were assessed. The searchers were provided with definitions of relevant, partially relevant and non-relevant which had their origins in the ISILT (Information Science Index language test) project undertaken by Keen and Digger in the early 1970s (Keen, 1973 ,) and which have been used in various tests since that time. Since the searches were carried out on three different engines it is highly likely that identical items would be retrieved with a possibility that ?already seen? items are consciously or unconsciously judged to be less relevant the second or third time around. This is partly resolved in the varying order of the engines to which searchers were assigned. In this instance, however the impact was not considered to be great as the aim was not to evaluate the performance of the individual engines per se but to obtain a user?s expression of satisfaction with precision across the engines? results. Participants were then asked to rate on a five point scale their **satisfaction with the precision** of the search results. An overall rating of effectiveness was obtained by the users' assessment of the ***overall success*** of the search engine in retrieving items relevant to the information problem or purpose on a five point scale. To measure utility users were asked to rate on a five point scale the **worth of their participation**, with respect to the resulting information; the contribution the information made to **the resolution of the problem; satisfaction with results; the quality of the results**; and the **value of the search results as a whole**.

Participants were asked to rate on a five point scale the ***overall success*** of the search engine in terms of the actual usefulness of the items retrieved. To measure interaction the users were asked to rate on a five point scale satisfaction with **query input, query modification, query visualisation, manipulation of output** and **satisfaction with visualisation of representation** of item. To measure efficiency participants were required to record the **search session time** and to rate on a five point scale the ***overall success*** of the search engine in retrieving items efficiently.

Analysis and interpretation

Our primary aim, expressed in proposition one, was to explore the multidimensional nature of users' evaluation of a system. To begin to understand how users might themselves evaluate these systems we correlated users' overall success rating with the ratings assigned on the four criteria to suggest which, if any, appears be the most important or contributory factor to users' overall judgement of system success. The results are shown in Table 2: where a moderate correlation is defined as greater than 0.4 and less than 0.7 and a strong correlation as between 0.7 and 0.9.

	Spearman's rank correlation coefficient **=strong *=moderate strength correlation			
Criterion	**Global**	**Excite**	**NorthernLight**	**Hotbot**
Effectiveness	.759**	.779**	.795**	.729**
Efficiency	.817**	.843**	.908**	.741**
Utility	.710**	.362	.930**	.806**
Interaction	.592*	.511*	.660*	.580*

Table 2: Global and engine level - Overall success rating correlated against the four criteria

The strength of the correlation ratings indicates that users' overall success rating of the system appears to be multidimensional. The Efficiency criterion held the strongest correlation with users' overall rating (.817) and Interaction the weakest (.592) which is in accordance with other studies which suggest ease of use and response time to be strong determinants of satisfaction. This result is fairly consistent with the analysis done at the level of the individual search engine, although we note the strong correlation with utility on NorthernLight and HotBot which was not found with Excite.

Indicators of the dimensions

Correlations were also taken of the measures within each criterion with users' overall rating. Strong correlations may suggest single measures as the best indicator of users' evaluation of the criterion. Results at the global level, across the three engines, suggest that 'satisfaction with precision' has the strongest correlation (.733) with the judgement of the systems' effectiveness. All the measures of utility held a strong correlation with its overall judgement, the strongest measure being 'satisfaction with results' (.755). The measure 'rate worth of participation' had a negative correlation which indicates that as utility rises the value of participation decreases. This may bring into question users' interpretation of value of participation. We can only speculate but it was possibly mistaken to refer to amount of user effort exerted. Interestingly, the measure of 'time taken to search' held a negligible correlation (.062) with the judgement of the systems' efficiency. The judgement of the satisfaction with systems' interaction held a correlation of moderate strength (.506) with the measure 'satisfaction with query visualization'. This is followed by 'satisfaction with facility to input query' (.486), 'satisfaction with visualization of item representation' to ease understanding of item/s from the hitlist (.452), 'ability to modify query' (.437), and 'ability to manipulate output' (.132), this being a very weak correlation. At the level of individual search engines, the judgement of interaction on both Excite and HotBot held the strongest correlation with satisfaction with query visualization, whereas NorthernLight held the strongest correlation with the ability to modify the query. The low correlations of the measure of efficiency and the negative correlation of the measure of value of participation are considered in the concluding section of this paper.

User derived reasons for attributing satisfaction ratings, overall and on each criterion, were collected using open-ended questions. These were analysed to suggest the extent to which our selected measures compared to those which users themselves might base an evaluation of system success. Some 250 comments were collected and, in the main, simply confirm the users' understanding of our intended interpretation of system effectiveness and utility. The selected, but representative, comments in **Table 3** show the user-derived reasons for assigning ratings of success of effectiveness seem to confirm the finding that user satisfaction with precision held the strongest correlation with user ratings of effectiveness. The utility measures all held strong correlations and the user-derived reasons also indicate that these were measures which users may themselves use. Further analysis would be required to ascertain if in fact these measures were simply variations of the same measure 'satisfaction with results'.

Effectiveness 'Information retrieved was extremely relevant to my needs.' 'Most items retrieved appear to have some relevance.' 'Too much irrelevant information.' **Efficiency** 'Ease of use.' 'Had to redefine search twice.' 'The search terms were attempting to pin down a concept that was hard to verbalize [or] encapsulate.' 'Would become "extremely efficient" as the user becomes more adept with search terminology phrasing and when an "advanced search" would be more appropriate. 'Very quick, only had to search once.' 'One search term located all items that were of some relevance.' 'Needed to define search better.' 'Minimum effort, but results not good.' 'Search engine seemed efficient enough, but the search term was unusual. I think with a more concrete search term the SE would have performed well.'	**Utility** 'Those items were found useful.' 'I have gained further info[rmation] on the subject I was search[ing] for.' 'Current info[rmation] was located.' **Interaction** 'I changed the query once and it was helpful.' 'The SE easily allowed the query to be modified.' 'Found it hard to refine search.' 'The query was easy to change but yielded no better results.' 'Good options to change query.' 'Refining the search was hard, I couldn't think of any new queries and the SE didn't offer any help trying to narrow down search queries, like the SE I usually use.' 'Could lead to different routes of enquiry from the initial search term.'

Table 3: User derived reasons for assigning success ratings on the four criteria.

The correlations found with the interaction measures were relatively low with satisfaction with query visualization holding the strongest correlation. The user derived reasons, however, indicate that there was perhaps some expectation that the system would provide some assistance in modifying the query and that this would impact on evaluation of system interaction. The efficiency criterion held the strongest correlation with an overall judgment of success, yet the measure of search time held a low correlation as a measure of efficiency. Interestingly, the user-derived reasons for assigning ratings of success on this criterion suggest that users relate efficiency, not to time taken but to something equating to the amount of user effort required to conduct a search. The real worth of this investigation is that while we may suggest that user evaluation is multidimensional and there are some obvious candidate variables on which to base the measure of these evaluation dimensions, considerable research is still needed to properly ascertain and validate user measures. We still need to understand how users themselves evaluate a system and to what these measures relate.

Feature impact

Our second proposition is that system characteristics will affect user evaluations on the task dimensions which have defined the evaluation. Users' ratings varied across the search engines and, within the scope of a small scale study, this was viewed as indicative that user evaluations are not random but an elicited response reflecting the support of the system to the users' task. As has been noted, strong correlations with users' overall rating and utility was found on NorthernLight and HotBot, in contrast to the very weak correlation found on Excite. The marked difference in the strength of correlation found between the systems is interesting but only in that it **suggests** that users' overall judgement may be more strongly associated with a judgement made on a particular criterion depending on the system. In a full scale evaluation study with a far larger sample more insight and interpretation could be possible from an analysis of the central tendency on the rating scales for the individual measures. For example, taking the strongest correlation with users' overall rating and interaction on NorthernLight (see Table 2), it was found that 77% of users of this system rated the ability to modify the query whereas globally and on the other two systems the measure of user satisfaction with query visualization had the strongest correlation . Again this may indicate the influence of a system feature on the users' judgement, but this cannot be substantiated without a much larger investigation.

Impact of query context

Our final proposition was that users' evaluations of the system may be moderated by some contextual characterisation of the user and information query. To obtain some indication of the support for this proposition we analysed the user/query context where a system received high/low ratings by correlating the four task identifiers (task defined, task purpose, task knowledge and task probability) against the overall satisfaction rating and the four criteria. Again we stress the exploratory nature of this exercise and that a greater sample would be required in an evaluation situation to support any analysis at this level.

Globally, across the three engines, moderate strength correlations indicated that as *task definition* increases so does overall rating of system success (.407), and satisfaction with effectiveness (.418) and efficiency (.482). The correlations between *task definition* and utility (.307) and interaction (.221) were weak. Weak or very weak correlations were obtained between *task purpose, task knowledge,* and *task probability* and the overall success rating and the four criteria. The suggestion that a system receives a higher rating of effectiveness and efficiency when the user has a well-defined task is not surprising. It would be reasonable to assume that in such a context the information seeker will have a fairly good idea of the search requirements to obtain good results. Indeed, the effect of the moderating context will be of more interest if notable variations can be found between the systems evaluated.

Indicative of such a finding is with the system NorthernLight. Whilst weak correlations were found globally for *task purpose*, when based on the data obtained for NorthernLight moderate strength correlations were found with Efficiency (.636) and Utility (.577). The comparison is with the weak correlation found on Excite data with Efficiency (.161) and Utility (.002). It would not be surprising if a correlation was found for intent of task purpose and utility across all three engines: a broad query, open to many avenues, could lead to a high rating of the utility of the results. That such a finding is strongly held only on one engine could lead to speculation that a feature of the engine leads to results which better support a broad query. For example, NorthernLight boosts features related to the visual organisation and representation of the search results which may better support the user with a broad query. Indeed users of the system expressed a high level of satisfaction with item visualization, in terms of understanding the content of the hit list, and query visualization in terms of understanding the impact of the query on the results obtained. A larger scale study would be necessary to ascertain the significance of this rating when compared to others.

Conclusions

Our preliminary investigation into user evaluations of internet search engines would seem to indicate that these are determined by many factors which together may represent dimensions of some overall user judgment of the system. To explore the value of a framework for evaluation based on multi-dimensions we defined and grouped possible indicators of success on the search task process in which the user is engaged. This would seem to be a reasonable approach to the substitution of usability measures for the evaluation of interactive system objectives to maximize search for information retrieval. The success indicators of user satisfaction were drawn from existing measures or by their definition provided by the consideration of the system feature objectives. However, there is a need to ascertain from users themselves what these success indicators might be and to validate these as relating to task dimensions. This is clear, not least, from our finding of some discrepancy in how users might define system efficiency with user effort being a key influence on the user judgment of the system. Further research is thus required to develop the multidimensional construct of user evaluation as possibly a function of the system support for the user retrieval task process.

Ultimately our aim in developing the framework is to have user evaluations link to system features, thus allowing a system to score high in particular aspects but not necessarily in all aspects. In this study, we can only speculate with caution that a system feature of query modification contributed to the users' evaluations, and that there was some observed effect of task purpose as a moderating variable in one system. While this tempts comment, we can only speculate that this was attributed to a system feature which better supported a particular query context. Only in a full-scale evaluation could this be tested using appropriate statistical techniques, such as regression analysis, to express user judgment of the system as a function of the task defined indicators of success and to explore, within and across systems, the relationship held among the dependent and moderating variables. Obviously this would be an expensive undertaking, but one which would not only allow variation to be found in users' assessments across systems influenced by system features, but also variations within search dimensions as influenced by user query task contexts.

Acknowledgements

This paper reports on results of a feasibility study of user evaluation of internet search engines, Devise (Johnson *et al.*, 2001). Thanks goes to Re:source for its funding of Devise under grant number NRG/085 from January 2000-2001. The full report is available from our website http://www.mmu.ac.uk/h-ss/dic/cerlim/ and can be requested from the thesis and reports lending division of the British Library. This research is being developed in Department of Information & Communications at MMU, and thanks to Sarah Crudge for many interesting discussions on its development towards the registration of her PhD on user evaluations of search engines.

References

- Back, J. (2000). An evaluation of relevancy ranking techniques used by Internet search engines. *Library and information research news*, **24**(77), 30-34.
- Baeza-Yates,R. and Ribeiro-Neto, B. (1999). *Modern information retrieval* Reading, MA: Addison-Wesley.
- Belkin, N.J., Cool, C., Kelly, D., Lin, S.J., Park, S.Y., Perez-Carballo, J., Sikora, C. (2001). Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval.? *Information processing and management*, **37**(3), 403-434.
- Beaulieu, M., Robertson, S., and Rasmussen, E. (1996). Evaluating interactive systems in TREC. *Journal of American Society for Information Science*, **47(**1). 85-94.
- Belkin, N.J. and Vickery, A. (1985). *Interaction in information systems: a review of research from document retrieval to knowledge-based system*. London: the British Library.
- Boyce, B.R., Meadow, C.T., and Kraft, D.H. (1994). *The measurement of information science*. Reading, MA: Academic Press.
- Brajnik, G. (1999). Information seeking as explorative learning, in: Proceedings of the MIRA ?99 conference.? In: *the electronic workshop in computing series*.? [Available at http://www.ewic.org.uk]
- Brajnik, G. (1999). Information seeking as explorative learning. In: S. W. Draper, M. D. Dunlop, I. Ruthven, and C.J. Van Rijsbergen, *eds.*, *Proceedings of Mira 99: Evaluating Interactive Information Retrieval, Glasgow, April 1999*. British Computer Society, Electronic Workshops in Computing. Retrieved 17 July 2003 from http://www1.bcs.org.uk/DocsRepository/02800/2836/brajnik.pdf
- Chu, H. and Rosenthal, M.?Search engines for the world wide web: A comparative study and evaluation methodology." In: *ASIS ?96: Proceedings of the 59th ASIS annual meeting*, 33, pp.127-135. Medford, NJ: Information Today. Retrieved 8 July 2003 from http://www.asis.org/annual-96/ElectronicProceedings/chu.html
- Cleverdon, Cyril W. (1991). The significance of the Cranfield tests on indexing languages. In: *Proceedings of the 14th International Conference on Research and Development in Information Retrieval (ACM SIGIR ?91)*, pp. 3-12. New York:ACM Press.
- Cooper, W.S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, **24**, 87-100.
- Ding, W.and Marchionini, G. (1996). A comparative study of Web search service performance. In: Hardin, S. *ed.*, *Proceedings of the 59th Annual Meeting of the American Society for Information Science* Vol 33. (pp. 136-142.) Baltimore, MD: Information Today.

- Dong, X., and Su, L. (1997). A comparative study of Web search service performance. In: C. Schwartz and M. Rorvig, *eds. Proceedings of the 60th Annual Meeting of the American Society for Information Science.* Vol 34. (pp. 136-142). Medford, NJ: Information Today.
- Dunlop, M. (2000). Reflections on Mira: interaction evaluation in information retrieval.? *Journal of the American Society for Information Science*, **51**(14), 1269-1274.
- Feldman, S. (1998). Web search services in 1998: trends and challenges. *Searcher*, **6**(6), 29-39. Retrieved 8 July 2003 from www.infotoday.com/searcher/jun98/story2.htm
- Feldman, S. (1999). Search engines: the 1999 conference. *Information Today*, **16**(6). Retrieved 8 July 2003 from http://www.infotoday.com/IT/jun99/feldman.htm
- Fowkes, H. and Beaulieu, M. (2000). Interactive searching behaviour: Okapi experiment for Trec-8. *Paper presented at the British Computer Society Information Retrieval Special Group 22nd Annual Colloquium on Information Retrieval Research, Cambridge, 5-7 April.* Retrieved 17 July from http://irsg.eu.org/irsg2000online/papers/fowkes.htm
- Gauch, S. and Wang, G. (1996). Information fusion with ProFusion. In: H. Maurer, *ed. Proceedings of the World Conference of the Web Society (Webnet ?96), San Francisco, CA, Oct 15-19.* Retrieved 8th July from http://www.csbs.utsa.edu:80/info/webnet96/html/155.htm.
- Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. *Information Processing and Management*, **32**(1).11-18.
- Harman, D. (2000). What we have learned and have not learned from TREC. *Paper presented at the British Computer Society Information Retrieval Special Group 22nd Annual Colloquium on Information Retrieval Research*, Cambridge, 5-7 April.
- Harter, Stephen, P. and Hert, C.A. (1997). Evaluation of information retrieval systems: approaches, issues, and methods. *Annual Review of information Science and Technology*, **32**, 3-94.
- Hawking, D., Craswell, N., Thistlewaite, P., Harman, D. (1999). Results and challenges in web search evaluation. In: *The Eighth International World Wide Web Conference*, Toronto, May 11-14. Retrieved 8 July 2003 from http://www8.org/w8-papers/2c-search-discover/results/results.html
- Hawking, D., Craswell, N., Bailey, P., Griffiths, K. (2001). Measuring search engine quality. *Information Retrieval*, **4**. 33-59.
- Hersh, W., Over, P. (2001). TREC ?9 Interactive Track Report.? In: *The Ninth Text Retrieval Conference (TREC ?9)*. Gaithersburg, MD: National Institute for Standards and Technology. Retrieved 8 July 2003 from http://trec.nist.gov/pubs/trec9/t9_proceedings.html
- Hildreth, C.R. (2001). Accounting for users? inflated assessments of on-line catalogue search performance and usefulness: an experimental study. *Information Research*,**6**(2). Retrieved 8 July 2003 from http://informationr.net/ir/6-2/paper101.html.
- Johnson, F. C., Griffiths, J.R., and Hartley, R.J. (2001). *Devise: a framework for the evaluation of Internet search engines.* London: Resource: the Council for Museums, Archives and Libraries. (Library and Information Commission Research Report 100)
- Keen, E.M. and Digger, J.A. (1972) *Report of an information science index language test.* Aberystwyth: College of Librarianship Wales.
- Keen, E.M. (1973) The Aberystwyth index language tests. *Journal of Documentation* , 29(1), 1-35.
- Leighton, H.V. and Srivastava, J. (1999). First 20 precision among World Wide Web search services (search engines). *Journal of the American Society for Information Science*, **50**(10), 870-881.
- Nahl, D. (1998). Ethnography of novices? first use of Web search engines: affective control in cognitive processing. *Internet Reference Services Quarterly*, **32**(2), 69.
- Robertson, S.E. and Hancock-Beaulieu, M. (1992). On the evaluation of IR systems. *Information Processing and Management*, **28**(4), 457-466.
- Salton, G. (1989). *Automatic text processing: the transformation, analysis and retrieval of information by computer*. Reading, MA: Addison-Wesley.
- Sandore, B. (1990). Online searching: what measure satisfaction? *Library and Information Science Research***,** **12**, 33-54.
- Saracevic, T., Kantor, P., Chamis, A.Y., and Tirvison, D. (1988) A study of information seeking and retrieving. I Background and methodology.? *Journal of the American Society for Information Science*, **39**(3), 161-176.
- Saracevic, T. and 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.
- Sherman, C. (2000). *The FireworksFly* [Available at websearch.about.com/library/weekly/ aa041800b.htm]
- Sherman, C. (2000). *'Old economy' information retrieval clashes with 'new economy' Web upstarts at the Fifth*

Annual Search Engine Conference: Conference Report. Medford, NJ: Information Today. Retrieved 17 July 2003 from http://www.infotoday.com/newsbreaks/nb000424-2.htm

- Stobart, S. and Kerridge, S. (1996). An investigation into World Wide Web search engine use from within the UK ?preliminary findings *Ariadne*, No. 6. Retrieved 8 July 2003 from http://www.ariadne.ac.uk/issue6/survey/
- Su, L. (1992). Evaluation measures for interactive information retrieval. *Information Processing and Management*, **28**(4), 503-516.
- Su, L. (1998). Value of search results as a whole as the best single measure of information retrieval performance.?*Information Processing and Management*, **34(**5), 57-579.
- Sullivan, D. (2000). Web search engine trends and achievements since the 1999 Boston Search Engine meeting, In: *Search Engines Today and the New Frontier: the Fifth Search Engine Meeting.*, Boston, Massachusetts, April 2000. Retrieved 8 July 2003 from http://www.infonortics.com/searchengines/sh00/sullivan_files/frame.htm. [PowerPoint presentation]
- Tomaiuolo, N.G. and Packer, J.G. (1996). An analysis of Internet search engines: assessment of over 200 search queries. *Computers in Libraries*, **16**(6), 58-62.
- Voorhees, E. and Garofolo, J. (2000). The TREC spoken document retrieval track. *Bulletin of the American Society for Information Science*, **26**(5). Retrieved 17 July 2003 from http://www.asis.org/Bulletin/June-00/voorheesgarofolo.html
- White, R.W., Jose, J.M. and Ruthven, I. (2001). Comparing explicit and implicit feedback techniques for web retrieval: TREC-10 Interactive track report, in:*The tenth Text Retrieval Conference (TREC 2001), Gaithersburg, Maryland, November 13-16.* Gaithersburg, MD: National Institute of Standards and Technology. Retrieved 8 July 2003 from http://trec.nist.gov/pubs/trec10/t10_proceedings.html
- Wiggins, R. and Matthews, J.(1998). Plateaus, peaks and promises: the Infonortics ?98 search engine conference. *Searcher*, **6**(6). Retrieved 8 July 2003 from http://www.infotoday.com/searcher/jun98/story4.htm

Find other papers on this subject.

"search engines" "user evaluation" "information retrieval"

Scholar Search

"search engines" "user evaluation" "information retrieval"

Google Search

How to cite this paper:

Johnson, F.C., Griffiths, J.R. and Hartley, R.J. (2003) "Task dimensions of user evaluations of information retrieval systems"*Information Research*, **8**(4), paper no. 157 [Available at: http://informationr.net/ir/8-4/paper157.html]

Check for citations, using Google Scholar