Centre for Research in Library and Information Management

cerlim

the MANCHESTER METROPOLITAN UNIVERSITY

DEVISE
A framework for the evaluation of Internet search engines

by
F C Johnson, J R Griffiths and R J Hartley

Resource: The Council for Museums, Archives and Libraries
2001

Abstract

This project investigates the feasibility of the use of user-satisfaction as a multidimensional evaluative construct of search engines. Search engine developments are reviewed to reveal a range of indexing and retrieval techniques that may assist casual users in the information retrieval task. Yet few evaluation studies have considered the impact of system features, in particular those with which the user interacts for search assistance. A broad review of retrieval system evaluation highlights the complex environment in which measures of both the utility of the search results and the usability of the system are sought from a user-perspective. Our proposed approach for a user-centered evaluation is based on a conceptual framework in which user-satisfaction is characterised as a variable dependent on system features and functions and expressed in a moderating context of user-task requirement. Towards this end, the research reported here focuses on the definition of the construct of user satisfaction on the multi dimensions of the retrieval process, an expression of what a typical user is trying to do. Empirical work was then undertaken to test the feasibility and potential value of the implementation of the framework for the evaluation of three search engines. Initial results are presented which provide a degree of understanding of how users are satisfied and on what criteria. This provides the basis on which we make recommendations for the refinement of the multidimensional framework and its use as a methodology for the evaluation of search engines from a user perspective.

For further information or correspondence please contact the Project Director or the Project Research Fellow at the address below:

Dr Frances Johnson
Project Director

Jillian Griffiths
Research Fellow

Centre for Research in Library & Information Management (CERLIM)
Department of Information & Communications
The Manchester Metropolitan University, Geoffrey Manton Building
Manchester, M15 6LL
Tel: +161-247-6156; Fax: +161-247-6351
F.Johnson@mmu.ac.uk, j.r.griffiths@mmu.ac.uk

The opinions expressed in this report are those of the authors and not necessarily those of Resource: The Council for Museums, Archives and Libraries

Library and Information Commission Research Report 100

RE/085

ISBN 1 902394 61 5
ISSN 1466-2949

Library and Information Commission Research Reports are published by Resource: The Council for Museums, Archives and Libraries and may be purchased as photocopies or microfiche from the British Thesis Service, British Library Document Supply Centre, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, UK.

Content Page

List of Tables

Chapter 1. Introduction

The overall aim of this project is to develop a framework for the evaluation of Internet Search Engines with an emphasis on a user-centered perspective. Towards this end we adopt the perspective that user satisfaction is a complex and multidimensional construct which is determined by the user's task requirement. Measures based on the resulting criteria provides a conceptual framework for system evaluation in which user satisfaction is characterised as a function of system-task fit expressed in a moderating context of the user requirement. The evaluation framework was developed based on a theoretical understanding of previous approaches to evaluation and some empirical work was undertaken to test its feasibility. The main objective of this feasibility study thus was to understand how users are satisfied and on what criteria. By focusing the measures for each criterion on the features of the system designed to support users in retrieving information, use of the evaluation framework may provide system designers with further insight into areas for development. In addition, the incorporation of a moderating context of user and task in the framework, as a possible influence on user satisfaction with system performance, is intended to provide a better understanding of why a system can receive varying evaluations across differing contexts.

The structure of the report is as follows. ***Chapter 1*** provides a general introduction to search engines and their evaluation and provides the rationale for our stated aims and objectives. ***Chapter 2*** charts the development of search engines to highlight the major factors which may impact on their performance. General observations on search engine usage leads us to focus on the more novel features which concentrate on helping the user phrase more effective queries and navigate through the results displayed, those which, in other words, may help the inexperienced searcher get to the information requested. This categorisation of features, especially those with which the user interacts during the retrieval process, provides us with important clues towards an evaluation methodology. ***Chapter 3*** reviews methodologies for IR system evaluation ranging from a system perspective of evaluating performance to a user perspective of gauging satisfaction. The resulting broad definitions of the criteria on which evaluations are based structure our review of methodologies for the evaluation of search engines. The intention is not to provide a comprehensive review but to set the goal of evaluation. That is, we explore the why and what of evaluation, against the criteria on which the evaluation is to be based, which in turn defines the system and context parameters in the design. In ***Chapter 4*** we construct a framework for evaluation based on the dimensions of the information retrieval task and define the criteria on which we develop a set of user satisfaction measures with system-task support. Testing of the framework is described in a small-scale implementation and initial results presented. We conclude the report with recommendations for the refinement of the multidimensional framework and its use as a methodology for the evaluation of search engines from a user perspective.

1.1 Rationale for a user-evaluation of search engines

1.1.1 Internet Search Engines

Internet Search Engines have proliferated with the growth of the Internet itself. There are a growing number of major general purpose search engines from both established commercial firms in the industry and fromnew up-and-coming technology firms, often emerging from university departments. A small number of these may ultimately become the winners, but prior to reaching this status the current state of play is of development and competition (advances in search engine technology are reviewed annually at the Infonortic's Search Engine Conferences: Wiggins and Matthews, 1998; Wiley, 1998; Feldman, 1999; Sullivan, 2000).

Search engines are often categorised as robot-driven which respond to a user query or directory-based systems that guide users through classified lists. Whilst this distinction is increasingly blurred with catalogues and full text indexes coming together in a single service, the popularity of the query-based approach is evident. A recent survey commissioned by RealNames (Sullivan 2000a) revealed that 75% of frequent Internet users use query based search engines and 70% of those surveyed said they know specifically what they are searching for when they use a search engine. Given their popularity, in this preliminary investigation of a framework for evaluation we focus on query-based search engines. This decision does not preclude development of the framework for the evaluation of subject-based (or combined) services. However some adaptation would be required to re-define the criteria for evaluation within the framework to reflect the general browsing task which subject- based services support with associated features such as visualisations of the information space. Focusing on query-based search engines alone allows us to delimit a range of particular key features in defining the criteria for their evaluation.

In general, Internet search services are built using 'spiders' or software programs to create and maintain a proprietary index of web documents, and a search engine, the underlying technology for retrieval, and the interface for users' search specification. Search Engines exhibit a number of key characteristics which have enabled them to develop rapidly and gain popularity for accessing global networked information. They are fast, robust, scalable, sustainable and use a variety of techniques derived from 30 years of research in IR to achieve their performance levels. However, considerable variation exists between the engines – in the techniques used for indexing, ranking, the search features, and the display of retrieved results – all of which can affect performance. Indeed, in such a context, it is not surprising that each engine is developing characteristics which may allow it to stand out from the others. Excite touts its concept search capability targeting the consumer market; Infoseek with its emphasis on company information targets the business user; NorthernLight offers serious information searching and has received much attention in the literature with its custom folders offering a visual overview of a search (Feldman, 1998). Suggestions are still being made for the next generation of search engines, and it is in this context that we can state that it will be some time before the technology reaches its users' expectations for finding precise information.

However, as Evans (in Wiggins and Matthews, 1998) and others (such as, Larsen, 1997; Berghel, 1997) have noted search engine developers may be approaching a fundamental limit in terms of the capabilities of their systems. Evans describes an uncertainty principle, holding that IR systems cannot automatically accommodate all idiosyncratic viewpoints saying "the best we can expect is for systems to be tuned to the expectations of the masses, with rapid adaptability to a given individual's viewpoint" (p. 16). Feldman (1998) warns that the problem facing developers is more fundamental, stating that the "Web searching market is extremely fluid and undefined. Hard as it is to design a television set or a car that everyone will want, at least manufacturers of reasonably standard products know why people will want them and what they will do with them. The situation is much less certain in the online world, in fact it is downright murky" (p.3).

The uncertainty which surrounds users' expectations and usage of search engines gives rise to the question as to how we can evaluate the impact of their developments on performance. More specifically, it is critical that we have some means to measure the impact system features have on users' satisfaction with respect to what they want to do or achieve with these systems.

1.1.2 Evaluation

Evaluation is a process by which the effectiveness of a service or system is assessed, in particular to establish the degree to which the goals and objectives are accomplished (Harter & Hert, 1997). The general objective of an IR system is to retrieve relevant documents for a given query, whilst at the same time to minimise the user effort in locating needed information. Thus the evaluation of a retrieval system can be seen to encompass many different viewpoints, from the mechanical (does it retrieve relevant documents for a given query, including the impact of design such as the use of natural language or controlled language indexing?); to the human (does it provide useful, usable tools, and how should the interface be designed to simplify user-system interaction?); through to the utility perspective for a given group of clients (does it deliver the information in a convenient form, in a timely fashion?) (Large et al., 1999). Evaluation from a user perspective is so broad, however, that it must embrace all these viewpoints. Further given the situation described above, in which systems are designed to meet a spectrum of users, information needs, and search behaviours the impact of these evaluation views on user satisfaction is likely to vary considerably across different contexts.

A broad comparison of the criteria and measures of user satisfaction proposed for the evaluation of retrieval systems set against the possible system and external (or contextual) parameters illustrates the potential for a highly complex evaluation situation. This is done in Table 1 which compares the following researchers' criteria in the **corresponding typeface**:

- the six criteria for the evaluation of a retrieval system, as identified in Cleverdon (1978)
- the criteria for the evaluation of interactive retrieval systems from a user perspective, as identified in Su (1992)
- the criteria for an evaluation methodology of web search engines, as identified in Chu and Rosenthal (1996)
- the recent recommendations of criteria for the evaluation of search engines in Oppenheim, et al., (2000).

Whilst this is a broad comparison, it highlights a number of important points with respect to the aim of this project. First, these criteria and their measures have been consistently used in evaluations spanning four decades. Second, in proposing the criteria for the evaluation of search engines certain adjustments or indeed alternative measures are recommended. These are discussed in more detail in the review where we focus specifically on the difficulties which arise in validating the use of traditional recall and precision measures when computed from an Internet retrieval situation as distinct from the test conditions of their origin. Third, while relevance based measures dominate, other factors such as the utility of the retrieved results, and the user interface may affect user satisfaction and thus have an important role to play in users' selection of systems. Further, the table sets a range of system components and user/context parameters against each criterion to attempt to show the role of each. For example, the technology comprising indexing techniques and retrieval algorithms could impact on retrieval performance.

These parameters become increasingly complex as the measures become more user-oriented as not only do they define what is evaluated but equally the parameters impact on the measures for the criteria. It is obvious, for example, that the content of the database or index searched will partly determine the items retrieved, and thus impact on a user's perception of the usefulness of the service in meeting the objective to retrieve useful items. However, the user judgement of utility, based on the value of the retrieved items, is distinguished from the criteria of aboutness used in the relevance measures of recall and precision. A user's judgement of system success based on utility may be influenced by a number of user factors, such as the context of the query, and the psychological state of the user. Thus such a judgement could be partially determined by a range of system factors, such as the speed of operation, the quality assurance of results or the presentation of results. The evaluation of the usability and functionality of search engines likewise must involve the user in some investigation of the search process the system supports and the impact the system features have on search behaviour as well as the retrieval outcome. The effectiveness of retrieval is partially dependent on the searcher use of the search features to formulate a query statement facilitating its intended interpretation. For example, a lack of precision may be caused by a searcher's reluctance to expend effort in narrowing a search. Indeed, the interface (and non-retrieval devices) may affect the whole mode of interaction for the user and hence influence the demands the user indirectly puts on the back end search technology. A further indication of the layer of complexity added as we move from the more abstract performance measures to those which involve the user lies in the consideration that the characteristics of the users' tasks may also influence their search behaviour

Table 1 Comparison of evaluation criteria and system/context parameters

Evaluation Criteria[1]		**System parameters**	**User context**
1. **Coverage** (proportion of literature on a topic)	• Composition of web indexes • Coverage, using the Clarke and Willett method	Composition of the index will affect the performance of the search engine	
2. **Recall** (retrieve relevant items) **Precision** (hold back non-relevant Items) • RELEVANCE (precision, relative recall, user vs. system ranking)	• Retrieval performance based on precision • Performance based on precision and relative recall	The indexing language, exhaustivity and specificity, and retrieval mechanism will affect performance	Query formulation, and search strategy
4. **Response time** (from request to results) • EFFICIENCY (search session time, relevance assessment time, cost) • USER SATISFACTION with response time	• Response time • Response time	As above, and organisation of stored documents, size of collection, file format will affect response time	As above, and type of query
5. **Utility** (worth of search results, and value of search results as a whole) • USER SATISFACTION with search results (importance of completeness and precision of search results)	• Overall quality of results as rated by users, and consistency of results, proportion of dead or out-of-date links and duplicate links	As above	As above, and user/ information need context
6. **Format (**presentation of the search results) • USER SATISFACTION with output format	• Number of output options offered, and analysis of the content of the output. • Options for display of results, and length and readability of abstracts	Type of display of output will affect performance in an interactive system	As above, and specifically user ability to judge document relevancy
7. **User effort** (expended to achieve a satisfactory response • USER SATISFACTION with search interface and online documentation	• User effort based on analysis of documentation and interface • Evaluation of GUI for user friendliness, and helpfulness of help	Interface, facilities for interaction with system and guidance	As above, and specifically the usage of interactive functions and user search behaviour

[1] some are measures which may group or define the underlying criteria

A possible consequence of the complexity of such interrelations among system and contextual parameters is the use of satisfaction as an evaluation concept. The construct of user satisfaction used in system evaluation aims to achieve such a summary expression of users' perceptions based on the usefulness of a system. Its appeal lies in its use as a surrogate measure of system effectiveness where a system is deemed to be successful if users' evaluations along various scales of satisfaction are at a maximum. Research into user satisfaction and its relationship (as a dependent or independent variable) to system acceptance and actual use and behaviour is extensively covered in the information systems management literature (Gatian, 1994; Parasuraman, et al., 1985, 1988; Goodhue, 1995). Yet relatively little work has come from the IR community in the definition of a satisfaction construct and the validation of user satisfaction scales and surveys (Harter and Hert, 1997. p38). A possible reason is that, in the context of user information searching, how users themselves evaluate system performance may be on multiple dimensions. Thus an expression of satisfaction on which system evaluation from a user perspective is based is a complex construct determined not only by a range of system influences (both the performance output and mode of interaction) but also influenced by a range of user contexts and requirements. In 1977 Tessier et al. made the assumption that user's satisfaction will be a function of how well the product fits their requirement and is experienced within the framework of their expectations. While this implies how satisfaction should be measured, our aim is to develop these assumptions to maxims on which to base the development of a framework for the evaluation of search engines.

1.1.3 Development of a framework for user-centered evaluation of search engines

A conceptual framework for the evaluation of search engines from a user perspective is thus proposed which is based on the notion that user satisfaction with a retrieval system can be characterised as a function of system-task fit. To this end, we identify a general task model of the retrieval process which it is assumed all retrieval systems will aim to support. Each of the steps in the process model provides some statement of user-requirement, what the goal-directed user is trying to do with system, and suggest the dimensions of user satisfaction criteria. By linking the evaluation criteria of system effectiveness, efficiency, utility and interaction to the task dimensions the measures devised for each were identified from the system components or features which supported the user in their task. Finally user-context parameters were identified for the analysis of their impact as a moderating context for the evaluation framework.

To test the framework the following measures were used:

1. User satisfaction measures of system effectiveness, efficiency, utility, and interaction. Users were asked to rate the system based on degree to which the system supported them in the associated task dimensions
2. Users and their information tasks were characterised based on questions which captured the nature of the information query and users' intent, amount of prior knowledge, and expectations.

The data collected was analysed with a view to understanding user evaluations of system satisfaction and thus addressed the following propositions:

1. User satisfaction is expressed as a multidimensional construct based on user requirement (what the user is trying to do). Correlations of user assigned system ratings on various scales were analysed to find which dimensions and measures appear to be the most important in defining user satisfaction.
2. User satisfaction is a meaningful evaluation of system characteristics. Within the constraints of a feasibility study we analysed satisfaction ratings across search engines to speculate on a possible link between user satisfaction and the system features which support users in their tasks.
3. User satisfaction is expressed in the context of an individual’s information need. The impact of the contextual characterisation of user and information query was explored to find the extent to which the importance of given system features is determined by user requirement and thus lead to a difference in the system evaluations obtained.

Chapter 2. Internet Search Engines: Factors which affect performance

The categorisation of search engines components and features which may impact on performance follows a logical sequence considering the database collection, the index, and the user/system search features. The tables are derived from more comprehensive reviews of search engines found in Su & Chen, 1999; Notess, 2000; Feldman, 1998; and Sullivan's searchenginewatch. We focus only on four search engines, AltaVista, Excite, HotBot (Inktomi), and NorthernLight to provide an indication of the variation found. We acknowledge that these illustrations may have some inaccuracies given the changing state of search engines, however the tabulation of features is not intended to evaluate or compare rather to highlight the characterisation of engines by features which provides clues for the development of an evaluation methodology.

2.1 Search Service Coverage

While it is possible to submit a web page to a search service for inclusion in its database, most services will also acquire database information from web pages through the use of agents or robots. Sullivan's table shows a number of factors which may vary across the strategies used by robots for crawling. Depth of crawling refers to strategies used for following inter-document links – some will follow all/ some will sample. The use of frames, and image-maps if not supported by the engine will impede progress in crawling the web. Learn frequency and instant index refer to strategies used to update the database with new or changed information. Some 'learn frequency' to re-examine sites which change frequently. Instant index refers to the time delay in which trawled pages appear on the index. While the process of selecting and/or reviewing quality content is generally reserved for subject-specialised search services, some query based services also attempt to reduce the size of the database by establishing subsets of reviewed resources or most popular ones. Link popularity when used to determine pages included in the index establishes the popularity of a page through analysis of the number of links there are to it from other pages.

Table 2 Search engine coverage

Coverage	AltaVista	Excite	HotBot (Inktomi)	NorthernLight
Estimated size	30m	50m		
Deep Crawl	yes	no	yes	yes
Frames support	yes	no	no	yes
Image Maps	yes	no	no	yes
Learns Frequency	yes	no	no	no
Instant Index	yes	no	no	no
Link popularity	no	no	yes	
Coverage (content)	www	www & reviewed sites	www	www & special collections/journal articles

2.2 Indexing strategies

The list of indexed elements in the representation varies from service to service. The majority will index every word on the page, others index only frequently occurring words, or words occurring within certain mark-up tags, or only the first x number of words or lines of HTML files. Stopwords may of may not be applied, and, if applied, may include words of very high frequency such as "web". The use of metatags, traditionally used to improve a search by providing a common ground of indexing terminology, is seemingly discarded by search engines. Web site developers have reportedly mis-used metatags, for example repeating terms many times, in the attempt to have a page appear in the top 10 retrieved. HotBot (Inktomi) reportedly enhances its index with human intellectual representations of items. Some services offer a combination of catalogs (selected collections described and classified into a taxonomy) and large full-text collections. These vary in the extent of human involvement for their creation and maintenance, and the way in which the alternative search modes are offered to the user.

Table 3 Search engines indexed elements

Indexing	**AltaVista**	**Excite**	**HotBot (Inktomi)**	**NorthernLight**
Full text	yes	yes	yes	yes
Stopwords omitted/ not searched	yes	yes	yes	no
Meta descriptions	yes	yes		
Meta keywords	yes			
Comments			yes	
Subject Categories				Uses people to create categories

2.3 Search features (user control of search)

The graphical user interface,GUI, of a search engine provides system designers with a mechanism whereby the control for interaction is placed with either the system or the user. AltaVista, for example provides the options for simple querying, advanced query, and predefined category browsing. In the opening screen (typically the simple query mode) most search engine interfaces focus on supporting the users' information seeking activities of query formulation and results display, in albeit a somewhat limited fashion. Typically the user is presented with an input box and possibly some guidance as to how to enforce the processing of the query terms (match all/ match any/ treat as exact phrase/ include or exclude a term). Although the interface for simple query appears straightforward to use (enter keywords, click submit, receive hundreds of results), beginner or casual users may find it difficult to use because of unfamiliarity with methods for narrowing search terms to retrieve a manageable number of hits to examine. The typical array of more advanced search capabilities are shown in Table 4. The use of these, for example boolean, to specify query term relationships and truncation or case sensitivity to facilitate the interpretation of a term, assume considerable experience on behalf of the user with some guidance offered in the help files.

Table 4 Search engine search features

Search	**AltaVista**	**Excite**	**HotBot (Inktomi)**	**NorthernLight**
Boolean search	yes	yes	yes	Yes
Nested parenthesis	yes	yes	yes	Yes
Include/exclude (+ -)	yes	yes	yes	Yes
Default	OR	OR	AND	AND
Proximity/near/adjacency searching	within 10 words	concept search approximates this	no	Relevance ranking gives boost for nearness
Phrase search	yes	yes	yes	Yes
Stemming/truncation (permit or inhibit automatic stemming, or specify truncation at the terminal)	yes	no	no	Automatic search for plural and singular word forms
Case sensitivity (wholly, partially)	yes	no	for a person search	Will boost rank if capitals in results when used in query
Fielded search (e.g based on title text, site, url, link, host, domain, anchor, image)	yes	no	yes	Yes
Limit restrictions (e.g. based on date, language, subject, document type, industry, domain, etc)	yes	yes	yes	Yes

2.3.1 Users & usage of search features

Search engines offer an array of search features found in traditional online services. Yet whilst many of these features give a trained search intermediary optimal search performance, search engine users are likely to range from experts to casuals (Travis, 1998). Wiggins and Matthews (1998) in summarising the themes of the 1998 Infonortics conference highlighted the consensus which was the driving force behind many of the developments reported. Professional searchers may be adept at using Boolean to refine searches but novice users are likely to become perplexed and frustrated. Thus it makes sense that on most search engines users are offered statistical based searches first. These are designed to act on natural language descriptions of an information need and to return a list of approximate matches as well as precise matches with ranking taking care of the potential overload of often long lists of near hits. However, whilst use of the retrieval models offered by these statistically based ranking algorithms is touted for end or casual users their effective implementation makes considerable demands seemingly beyond the average user.

Surveys of web usage give some sense of what the average web searcher is doing and point to differences between web searches and queries with traditional IR systems. Observation of average web searcher (Spink et al, 1998; Ellis et al., 1998) point out that their ineffective use may be owing to the little understanding most users have as to how a search engine interprets a query. Few are aware when a search service defaults to AND or OR, and expect a search engine to automatically discriminate between single terms and phrases. Further, devices such as relevance feedback, seemingly conducive to end-user searching, works well if the user ranks ten or more items, when in reality users will only rank one or two items for feedback (Croft, 1995). Most significant is the finding from a study which looked at one million queries put to Excite that users will enter one or two search terms rather than a full informative summary of the information query (Jansen and Spink, 2000). This is possibly due to difficulty in selecting terms arising from the way in which users are reported to conduct a search. Koll (1993) explains that users provide few clues as to what they want as many users approach a search not knowing exactly what it is they are looking for. In adopting the – *I'll know it when I see it, or the unknown needle in a haystack* – approach to information seeking, users cannot be expected to formulate a precise query.

Larsen (1997) is of the opinion that current Internet search systems are prototypes and that their development will not focus solely on the refinement of IR techniques to zero in on the perfect retrieval set. Rather alternative techniques will evolve to meet the behaviour of average web searchers. In recent years of their development there has been a notable shift towards the introduction of search features which appear to respond to the ways in which users actually search with these systems. Beyond the level of mere statistical keyword matching developments utilise a variety of technology features to help users get the information they want, even if it is not what they asked for. Such developments center on the areas of search assistance or query formulation with subsequent user control in modifying the query and navigation of the results. The notion that improved interaction may be the key to obtaining better results is attractive in principle but diluted by a cautionary observation from Nick Lethaby of Verity Inc paraphrased in Andrews (1996) that "users don't want to interact with a search engine much beyond keying in a few words and letting it set out results" (p.42). Thus in the context of categorising the development of search features we distinguish those which provide searcher assistance and those which shift the control back to the system to provide the most likely relevant hits.

2.4 Search features (system control of query)

As it can be assumed that most users do not use advanced search features, or enter complex queries, or indeed want much to do with searching or interaction, search engines are trying to automate query formulation. That is shifting the burden of coming up with precise or extensive terminology from the user to the system. Some tweaking in this general direction has already been shown in Table 2.3 for example where NorthernLight will boost the ranking of retrieved items containing capitalised query terms. More elaborate are the notions of concept searching, the use of site popularity to improve the relevance ranking of results, and the creation of directories to help the user browse more productively.

2.4.1 Query expansion

Help in improving a user's query formulation may be provided by use of concept searching. The assumption here is that users will take a quick and simple approach to putting a query to a search engine and that automatic expansion of the query will improve the search expression. On a deeper level, concept processing of a search statement is to determine the probable intent of a search (e.g., Excite's ICE technology)

Automatic query expansion uses a system-generated thesaurus, more accurately described as a list of words statistically related by frequency of co-occurrence in documents. Thus a search engine may modify a query by adding those terms with a strong association or high coincidence in documents containing the initial query term(s). This often results in high recall typical of a thesaurus-based system and, since precision can be adversely affected, the search may be subsequently refined by allowing the user to select relevant items for a reiteration of the search. Excite's ICE technology (1999) reportedly works at a deeper level applying concept processing to determine the probable intent of the query. Whilst detailed operation of the technology is confidential, some clue to its working is found in a comparison to Latent Semantic Indexing which analyses, by correlations of related terms, separable contents (or concepts) of a document. Probability theory may also be employed in concept processing to look at ideas contained in text as the outcome of probabilities derived from the clustering of certain symbols. For example, if the symbol 'bar' clusters near certain other symbols in a passage, such as 'drink' 'bottles', then it is likely to refer to a room containing a counter across which refreshments are served rather than rod, a place at which a prisoner stands, or a European sea-fish. Furthermore if these clusters of symbols are present in a text, there is a good chance that it is about the said concept even if the word 'bar' is not actually present. As far as the user should be concerned, the outcome of such processing is that relevant items may be retrieved even if they fail to contain the original keywords of the search statement. This is quite a significant advance on keyword matching when one considers the various ways in which an information query may be expressed, each as likely as each other, but which often result in little or no overlap in the results obtained when put to the same search engine.

2.4.2 Query modification

Providing more <u>user control</u> during query re-iteration and re-formulation, Excite's search wizard and AltaVista's refine function present to the user suggested search terms which frequently occur in the items retrieved. Infoseek's automatic categorisation of documents by topics is likewise offered as a browsable suggestion of topics likely to be relevant to a given search. All of which may assist the user in narrowing a search and provide more precision in the search results.

Another technique in providing user control in the process of query modification is the relevance feedback option (e.g., 'More like this'). This is where conventional querying and browsing strategies have been integrated to allow users to specify a particular document and then browse from that document in order to build a request model. This results in an iterative process consisting of query modification and feedback placing a user in control of the interaction. The basic principle being that users control subsequent queries by assessing the relevance of documents which are then used to modify subsequent query formulation. The query may be reinstated using high frequency terms from identified relevant documents or the entire contents of the specified document may be used as the search parameters to locate similar documents. Again, as far as the user is concerned such a search function assists in the specification of the query at an appropriate level without placing too much burden on the user coming up with the terminology to be used. To an extent the searcher is assisted in transforming a perceived information need into a search formulation within the vocabulary and command constraints of the system.

2.4.3 Query visualisation

Where some form of automatic categorisation of documents by search engines takes place an additional functionality may be offered in the form of the visualisation of multidimensional information about search results. That is the creation of on-the-fly groupings of search results can aid browsing of the different themes or concepts within the search results. Such organisation of results into categories reduces the potential overload in the retrieval of 100s or 1000s of items and assists the users in judging the relevancy of the retrieved items. It also has a useful side effect of highlighting to the user the potential ambiguity of the original search terms (as has been noted, users often fail to provide the important contextual information of a query) and thus can be viewed as a query assistance. Excite's ICE technology recognises clusters of documents and from this can base the grouping of the search results. Most elegant is NorthernLight's dynamic custom folders (Zorn et al., 1999*)* based on their categorisation of documents in which documents are mapped to a classification system and tagged accordingly. Custom folders based on the search results set provides the user with a hierarchical overview of the major topics retrieved allowing the drilling down from the broad to the specific aiding the browsing of different themes or concepts within search results.

2.4.4 Popular queries

Search assistance can thus be provided in the form of query expansion, query modification or visualisations of the major topics resulting from the query. These all work towards the general improvement of a typical search in which the user submits a couple of keywords, a strategy which eludes the capture of important contextual information of the need and specification of relationships among query terms. Most traditional information retrieval techniques rarely deal with a further complexity in the way in which humans are accustomed to conveying the meaning of or understanding discourse. Much of what we convey is in what is not said (as is what is said) when assumed by the context in which the query is stated. A user who enters the term 'penguin' to a search engine is probably searching for information on the bird rather than information on penguin books or the US rugby club. Similarly the user who enters the broad term 'travel' is probably looking for good travel reviews or pricing information on holidays, and would be less interested in the technical details of Stevenson's Rocket. Using a bayesian (probabilistic) approach to retrieval where knowledge of past events can be used to predict outcomes, prior knowledge of what users are searching for can be factored into the retrieval strategies of search engines.

AltaVista's "**Ask AltaVista**" is a version of the **AskJeeves** service. AskJeeves works on a large human generated database of questions based on what people actually search for. When a broad term is entered AskJeeves suggests a set of questions which the user may have intended or suggests a set of alternative, more specific queries. A more specific variation of this is AltaVista's **real names link** which will direct a user to official sites when a brand name search is conducted. HotBot's **related searches** offers searches which are similar, either more general or more specific, to a given query. Excite's **target results** responds to certain types of popular queries with targeted information at the top of its results pages. For example a search on a geographical location such as "New York" will offer first its list of pre-programmed results or custom information including a city map, tourism resources, current weather etc. In a sense the search engine infers that this is the type of information the user is most likely to be searching for when entering a general query.

Table 5 Search engine search features (system control)

Search features (system control)	**AltaVista**	**Excite**	**HotBot (Inktomi)**	**NorthernLight**
Query expansion		Concept search		Concept processing?
Query modification	Refine (suggest terms)	Search Wizard (suggest terms) More like this (browse feedback)		
Query Visualisation		Cluster/group search results		Custom folders
Popular queries	RealNames Related searches Ask altavista	Related searches Target results	Related searches	

2.5 Results display

Once a search is completed, display and browsing capabilities can help a user to determine which items are of interest. Most search engines will present the retrieved items 10 to a page in a default format showing at least the title and some text. Format displays can usually be changed with options such as: Sort by Date, Clustering by site/sort by URL (to identify pages from the same site and thus preventing any one site from dominating the results). The summary may vary in size and preparation, e.g., some are pre-prepared, automatically constructed, using text extracted from heading tags, first x words of text, or most frequent words. Where search terms are highlighted in the text, the user may gain some indication of why an item was retrieved and whether the context of the retrieved record matches the information need.

Table 6 Search engines results display

Display	AltaVista	Excite	HotBot (Inktomi)	NorthernLight
Sort by options	no	yes	no, but offers clustering	yes
Results at time	10	10	10	10
Title size	78	70	80	80
Summary size	150	395	170-250	150-200
Metatags description	yes	yes	yes	no
Highlight search terms				

2.5.1 Ranking

In terms of judging the results list Courtois and Barry (1999) argue that users are most likely to scan their results list and retrieve only selected items. However Cullis (in Sullivan, 1998) found that only 7% of users really go beyond the first three pages of results. Sullivan goes further saying "most users will find a result they like in the top ten. Being listed 11 or beyond means that many people may miss your web site " (2000). This suggests that users are rarely interested in a comprehensive, high recall search, but rather are satisfied with the retrieval of a couple of relevant hits.

Courtois and Barry (1999) point out the popularity of search engines is due in part from the perceived ease of use caused by their use of ranked output. The results and their relevancy to a given query are usually ranked by statistical term frequency, location, and possibly proximity of terms in the documents Simply put, a page which makes frequent mention of terms will get a higher rank than a page with only one reference. Similarly, a page with the search term in its title will be considered more relevant than others. How these criteria are applied defines the ranking algorithm and varies among search engines.

Hotbot describes term frequency and location as primary factors (Sullivan 1999a). Documents with more occurrences of the search term receive a higher weight, but the overall obscurity of the term within the database also has an impact. In addition, the number of occurrences relative to the document length is considered and shorter documents are ranked higher than longer documents with the same number of occurrences. Terms in the title or metatags are weighted higher than terms only within the text. AltaVista considers these factors, as well as the number of terms matched and the proximity of the search terms (AV Search: question 1999). Others provide less information. However Sullivan (1999b) reports that Excite does index terms in metatags, and retrieves documents by analysis of the document content for related phrases in a process it calls Intelligent Concept Extraction (Excite, 1999).

These methods for ranking output on predicted relevance have been experimented with for decades, but are limited to relevance based on topic alone. Barry and Schamber (1998) list at least a dozen further indicators which may determine the relevance of an item to a given user, including factors such as novelty, source characteristics, and availability. Given the utility of ranking, from a user point of view, in minimising the effort in finding an item, search engines have adopted a variety of experimental approaches using off-the-page parameters to boost the ranking of an item.

Link popularity boosts the ranking of a site if it is deemed to be popular based on the frequency with which other documents link to it. Generally speaking, counting links will set those with most pointing to it higher in the ranking. However, in practice the technology may be more complex whereby, for example, a link from a reviewed site or one with a good reputation will carry more weight in the overall analysis. Search engines using link popularity, such as Google, can be said to automatically capitalise on the human endorsements of web pages made by site authors when linking or pointing to what is in a sense recommended sites. A variation of this use of collective judgements is the use that can be made of the search behaviour of millions of web users in ranking popular sites. Direct Hit is a company which works with search engines (e.g. HotBot) and monitors user clicks on search results (what pages they visit). Over time, a measure is obtained on the popularity of sites – those which are visited more than others rise higher in the popularity rankings. To use this information in a search engine, the user may be offered the Direct Hit option on a page of search results. This will bring up the list of search hits ranked to be popular by Direct Hit. For example, in HotBot Direct Hit results are displayed under the heading “Web matches: top 10”. This is usually available only when a popular query is entered, and is usually most effective for one or two word queries looking for information on, for example, a famous person or a particular site. As a result the ranking of the results delivered by the Inktomi engine begin on the second page of ranked results.

Reviewed status gives pages a boost if a site is listed in an associated directory or forms part of the “reviewed” content provided by the search service. Meta-tags gives boost if a search term appears in a metatag.

Table 7 Search engines ranking boost

Ranking boost	AltaVista	Excite	HotBot (Inktomi)	NorthernLight
Link popularity	yes	yes	yes	yes
Direct Hit	no	no	yes	no
Reviewed status	no	no	no	no
Meta-tags	no	no	yes	no

2.6 Chapter Summary

The review has presented a very broad categorisation of search engine components to show the extent of variation of features offered by individual search engines which may impact on their performance.
Any combination of which may lead to a more effective search, and thus improved performance and ultimately user satisfaction with the retrieved results. In the context in which search engines operate (notably casual users) there has been an increasing trend to provide a range of search assistance features. Such that it could be argued, as in our introduction, search engine developers are targeting a niche, a type of user and/or information query. Future development is uncertain. Trends can be identified, such as automatic categorisation, information visualisation, and the use of bibliometrics on the web. The former may assist a user in understanding content of large collections or search results, the latter used to recommend documents by analysis of citation paths or hyperlink paths. It would appear that the shift towards supporting a user in their information seeking task, possibly to the extent of providing the information even if it was not requested, will continue to drive the advancements in techniques and technologies.

The problem faced by designers is that given the wide range of potential users little is known as to what users want, and how they might use these systems. Critically it is not known how users are satisfied and what impact these more novel features might have on search satisfaction. Thus, it is towards this end that we develop a framework for evaluation which asks **how** users are satisfied (e.g., whether it be on the retrieved results alone, and/or on their interaction with the system and assistance provided). Further, the framework incorporates a given spectrum of information needs and user types so that we can begin to understand the moderating effect of context on user-system satisfaction.

Chapter 3. A review of IR system evaluation

The focus of this section is on how the methods and past studies of IR evaluation can shape our understanding of what has been or can be achieved in evaluation of search engines. It is not and cannot be an attempt to review the approaches, issues, methods, underlying assumptions, findings or results of the many valuable evaluation studies of IR. Rather we broadly categorise the criteria on which evaluations are based to obtain an in-depth understanding of what is evaluated, the goal of evaluation, and the implications of evaluation in a complex situation with many interrelated system and context parameters. The section concludes with an insight into the shortcomings of the use of a user satisfaction construct as a surrogate measure of system success which leads us to consider the alternative perspective as conceptualised in the proposed framework.

3.1 Cranfield studies

Evaluation of comparative systems has a long tradition of improving the state of the art in IR technology. Researchers and system developers would like to test the truth of their theories about IR and/or to demonstrate a marked improvement in retrieval performance. The criterion for the evaluation of performance effectiveness has, in the main, been based on the overall goal of a retrieval system, to retrieve relevant documents. Such evaluations adopt the Cranfield-experimental model based on relevance, a value judgement on retrieved items, to calculate recall (or its surrogate, relative recall) and precision. These dual measures are then presented together where recall is a measure of effectiveness in retrieving all the sought information in a database, and precision assesses the accuracy of the search. In an experimental environment all variables can be controlled except those independent variables of interest (such as the indexing language) and comparisons based on these measures of effectiveness.

Many and various criticisms, problems and concerns have been leveled at the validity and reliability of a Cranfield approach to IR evaluation (e.g., Ellis, 1984). Most fundamental is the compromise necessary in defining relevance for such experimentation. For example, it is necessary to assume that relevance judgements can be made independently, that is a user will read each document as if new, without being affected by what they have learnt through reading previous ones. Furthermore, the predefined concept of relevance judgements on which recall & precision measures are based makes the assumption that relevance can ignore the many situational and psychological variables which in the real world affect relevance, which as Large et al (1999) state is “in the eye of the beholder”. In an interdependent system there may be many manifestations of relevancy unique to an information need which, in turn, is unique to a particular individual (with inherent variation). To assess the validity of the Cranfield measures for IR evaluation requires an understanding of relevance. The extent of the measurement errors introduced by variations in relevance assessments and “missed” relevant documents is essentially unknown, but has been shown not to affect relative results in comparative tests over a number of queries (Lesk and Salton, 1968). Indeed, in the defense of the Cranfield approach in which test collections are used, with information queries and preserved judgement sets, extensive sampling can be used in order to 'offset these compromises' (Salton 1992*)*

As such, this long standing approach to IR evaluation using precision and recall computed from a large body of evaluation data has come to be known as the traditional or default model of retrieval testing. More recently, since 1992, the approach has been embodied in TREC (Text REtrieval Conferences, funded by NIST/(D)ARPA) where participants use a standard large-scale test collection (Harman 1995, 1996) and compare system performance using standard evaluation measures of precision, aspectual recall, elapsed time and search satisfaction

3.1.1 Cranfield-type evaluations of search engines

The majority of studies evaluating search engine performance *(*Ding and Marchionini, 1996; Gauch and Wang, 1996; Tomaiuuolo and Packer, 1996; Chu and Rosenthal, 1996; Clarke and Willett, 1997) are based on some notion of relevance and thus regarded as Cranfield designs (Harter & Hert, 1997). For example, Leighton and Srivastava (1999) evaluated five search engines based on precision on the first 20 results returned for 15 queries asked at a University library reference desk. 'Top 20' precision rates the services based on the percentage of relevant results within the first 20 returned and uses a variant that adds weight for ranking effectiveness. Overall AltaVista, Infoseek and Excite performed best. They speculate that this may in part be due to cleaner databases with low duplicate or dead links, and common features which allow the user to control the search, such as case sensitivity for capitals. Typically, however, when used to evaluate search engine performance these measures are not computed from extensive evaluation data under test conditions to consider a limited set of environment variables. The appropriateness of this approach for the evaluation of web search engines must be questioned with respect to the origins, purpose and assumptions made in the use of recall and precision measures.

3.1.1.1 Recall

A major limitation is that search engines and the web do not provide for the controlled environment. As a result a majority of the studies which use a Cranfield approach report performance based on the precision measure only. Leighton and Srivastava (1999) chose to base the performance measure on precision alone because they argue that in their study, involving undergraduates precision is more important to the user than recall**.** That is, searches tend to be exploratory rather than comprehensive. This is a highly debated topic and is touched on in the next section on utility. The calculation of precision, however, assumes that the database is partitioned into retrieved and not retrieved. This is not the case in ranked output, hence the calculation of precision at various cut off points. Further, in essence, true recall cannot be calculated for searches in a web space because the total number of items returned by a search engine is too great. Thus it is not possible to calculate the number of potentially relevant items for a given query in such a huge and dynamic database. Given the dual nature of these measures it would seem advisable to at least attempt to approximate recall by the pooled method pioneered by the TREC experiments. Clarke and Willett (1997) developed such a method for comparing the recall of three sets of searches conducted on the different indexed collections of AltaVista, Excite and Lycos. Relative recall (the proportion of relevant documents retrieved with one engine amongst all relevant documents found using all search engines and strategies) was calculated by checking how many of the relevant documents found by one were present in the coverage of the other search engines.

3.1.1.2 Dynamic database

In the design of any evaluation experiment or investigation it is important that steps are taken to avoid the introduction of bias favouring one service over another. In the web environment the dynamic state of the databases searched (as indexes are generated using autonomous search robots) presents a particular difficulty for comparative evaluation. Most of the studies undertaken do acknowledge this and stress the need to run the queries on all engines at the same time, or in the briefest possible time period. The intention being to prevent bias towards a later evaluation where an engine has been able to index/retrieve new pages or re-fresh its index.

3.1.1.3 Relevance criteria

A further difficulty in use of these performance measures in an operational environment is that a lack of standardisation of the criteria for the relevance judgements across the various studies makes any attempt for comparison virtually meaningless. Tomaiuuolo and Packer (1996) for instance, do not define their criteria for relevance; others, such as Chu and Rosenthal (1996), have used a three level scoring method of 1 for relevant, 0.5 for partially/somewhat relevant, and 0 for irrelevant. However, as Oppenheim et al (2000) point out many have developed their own schema for scoring these points. Clarke and Willett (1997), for example, assigned the score 0.5 if a page consisted of a series of pages which lead to one or more relevant pages and, 0 to sites which could not be opened ("file not found" error message) or because of excessively low response times. Duplicates sites were penalised and scored as 0 (as in Leighton and Srivastava, 1999), but mirror sites were scored as unique. Nasios et al (1998) assigned one of five marks to each hit categorised as A – a best possible result; B - fairly relevant that partially or superficially covered the query theme or contained a link to a A type page; C - an irrelevant hit; X - failed to retrieve a web page due to broken link or server error; and D indicated a duplicate hit. Humphries and Kelly (1997) used a five score system where 4 was assigned to an authoritative site; 3 to an informative; 2 to an uninformative; 1 to unrelated/ irrelevant; and 0 – error.

Leighton and Srivastava (1999) based their criteria on Mizzaro's (1997) framework of the concept of relevance which views the relevance relationship as three components. 1) topic relates the information resource, (the document or information contained within) to the subject area of need. 2) task relates the resource to what the user wants to do with the information; and 3) context relates to everything else, that is, what the user already knows, what reading level the resource is at, how much time and money the resource will cost etc. They, the researchers, then defined the criteria for relevance categories based on topic along with anticipated tasks or information needs that would be represented by the request for the topic. Whilst they did not employ actual users to evaluate the results, stating that no other study reviewed had done so (p874), they could be seen to attempt to encompass an element of end-user relevance criteria for the evaluation of an operational system.

3.1.1.4 Queries

A final limitation in adopting a Cranfield approach for comparative evaluation is the requirement that the query expressions are kept constant across the engines. In general this results in the use of query expressions in their most basic form. Statements in the majority of studies were entered with no use of search features such as operators, modifiers or quotes. This, it could be argued, is a realistic approximation of the type of searching done by most users. Jansen et al (1998) found that of over 50,000 searches performed by 18,000 Excite users less than 7% used AND, and that +/- and double quotes were used in fewer than 6% of searches. However, Leighton and Srivastava (1999) using only unstructured or natural language queries suggest that the choice of search expression was a weak point in the design. It could be argued that in adopting a precision measure of system performance an underlying model of a user/ searcher is assumed. As a result, the most that can be said is that for a given set of users the system performed at this level.

Oppenheim et al (2000) conclude that the idiosyncratic approaches adopted by evaluative studies of search engines based on Cranfield render these inconclusive. They suggest that further evaluation of performance should consider alternatives to recall and precision. For example, Expected Search Length (Cooper, 1968) can calculate 'cost' in the sense of the number of sites a user looks at before sufficient items are examined to satisfy the query. They also recommend for a measure of performance the Back and Summers (unpublished) method which involves asking users to categorise a percentage relevance score to each hit. Both recommendations, it is noted, involve end users directly in making some judgement on the retrieved hits. An alternative evaluation infrastructure to reliably perform repeatable experiments in the context of the www is use of the Web track, new to TREC 8. The track used a frozen snapshot of the web as its document collection, known as VLC2, representing approx. 18.5 million web pages and 10,000 queries from logs from AltaVista and Electric Monk SE2. Participants submitted the top 20 documents for all 10,000 queries from which 50 queries were selected to judge the retrieved top 20. Results verified the ability of these systems to handle large amounts of data.

3.2 User-oriented evaluation of interactive retrieval systems

Search engine developers' responses to the TREC results reported at the Infornortics 5th search engines meeting are documented in a review from Chris Sherman "*The Fireworks Fly*" (2000). In defense of the relatively poor performance levels reported, developers considered the basing of performance on binary relevance judgements to be a poor match of the systems' objectives. Whilst the Cranfield approach to evaluation gives an objective measure of performance, it is based on the assumptions made with regards to the definition of relevance (in a sense, a predefined output). Suggestions of search engines' objectives as offered in Sherman's report include speed of results, getting information from users, browsing categories, and promoting popular sites. As the representative from AltaVista stated, search engines increasingly differ from each other in significant ways because no one model (an analogy was made to car models) will satisfy all needs. The objections to TREC imply that search engine developers would adopt, in preference, a methodology which attempted to evaluate the functionality of such developments from a user perspective. Further, given that it is acknowleged that search engines are targeting market niches, possibly user groups or types of queries, such contextual information will be important in any evaluation undertaken.

Alternative approaches to the evaluation of IR systems which involve the end user address the well known shortcomings of Cranfield, specifically its predefined output and input which the measures assume. The Cranfield methodology generally excludes the user (with an information need) from making the relevance judgement for the basis of the measures, which may indeed be inappropriate or incomplete measures from a searcher's point of view. With the advent of end user searching and for the evaluation of operational systems it has been argued that actual users of the system should make the relevance judgements to obtain a more realistic assessment of system performance from a user point of view. Furthermore, the approach arguably treats the system as a black box (Robertson and Hancock-Beaulieu, 1992) in making an assumption that the retrieval situation will be static, that is a one-off (offline) retrieval situation, with limited, if any, consideration of interactive searching by end users. Harman (2000) commenting on TREC as a test-collection-based evaluation points out that as such what is measured is the initial set of results users would see after they input a query but before any interaction. Whilst this point of measurement is important, and some users are satisfied with this, Harman states that "the average precision measure has strong recall component. The recall performance will only be further improved by user interaction and appropriate new tools." Put another way, in a search conducted outside of these experimental conditions, a lack of precision may be owing to a searcher's reluctance to expend effort in narrowing a search. Recent systems, including web searching, support a dynamic model of IR permitting interactive searching. In this model, it cannot be assumed that some preliminary preparation of query has been done, to put a one-off well-formed query to system, but rather the user will undertake extensive query reformulation via direct interaction with the system. An interactive IR system is thus one in which users' goals and strategies change in responding to messages of the system, and as Robertson and Beaulieu (1992) state "the rise of the interactive system has made evaluation methodologies that leave the user outside the system less and less tenable".

3.2.1 Utility

The involvement of users in the relevance assessment for performance evaluation based on recall and precision presents the difficulty that users will bring to the assessment whatever subjective criteria they wish, which with respect to a genuine (rather than invented) query is dynamic and situated in a moment of time-space. Indeed, research into end-user criteria for relevance has revealed a wide range of factors, other than topic, which may be bought to bear on the judgement (Barry & Schamber, 1998). Such that it has been widely debated that a measure to gauge system effectiveness should be based on the utility, not topic relevance, to the user of the documents retrieved. In such a user-orientated evaluation of system performance the user, seeking utility of the documents retrieved, can be influenced by a number of user factors, such as the situation context of the query, the psychological state of the user (e.g., frustration level) and logistics (e.g., time available).

Significantly, for a user centered evaluation of system performance, proponents of a utility measure raise some doubts as to the compatibility of the assumption that systems should aim to high recall and precision performance (Cooper, 1973). In their review Harter and Hert (1997, p.15) point to research which support Cooper's Utility Theory suggesting that users are not interested in topicality, precision, and exhaustive high recall but in the usefulness of the documents retrieved. Cleverdon (1991) asserts that recall is rarely a user requirement in operational systems. Meadow (1986) suggested that users are unconcerned with precision, and Sandore (1990) found that precision did not correlate with user satisfaction. More recently Su (1992) in an empirical investigation sought a single measure of system success for the evaluation of interactive systems from a user-perspective. She justifies the requirement for an evaluation methodology for system comparison and choice which involves the end user and their information problems in realistic operational IR situations. To this end, she posed the question whether, by correlating twenty measures of retrieval performance with users' overall judgement of system success, a single best indicator of a successful performance could be found. Her correlation identified seven significant variables and based on the strength of the correlation she found 'value of search results as a whole' to be the best single measure. This measure of utility, distinguished from the criteria of aboutness used in a relevance measure, was based on the users' satisfaction with and value of the retrieved items as a whole with respect to the actual usefulness of the items to the information searcher[2].

[2] [there was however suggestion from further data analysis that in the minds of users when assigning success ratings to a retrieval system completeness and value of search results may have been measuring the same thing. This would imply the importance of recall.]

The arguments for testing the performance of operational systems based on user judgement of output relevance or utility are strong, but add a layer of complexity to the evaluation methodology. Recall and precision measures can be applied to demonstrate, for example, the effectiveness of techniques for stemming in retrieval systems. The implications, in terms of understanding the influence of system components, on performance based on utility measures require careful consideration and interpretation. A range of system factors, such as those offered by search engine developers (see above, such as speed of operation, quality assurance of results to presentation of results) could impact on a user's judgement of system success based on utility. For example, a user's judgement of the value of the results may be partially determined by how novel the results are. In such an instance the order of presentation of the results is likely to impact on this judgement. Equally, a user may be influenced by the speed at which he/she is able to identify useful documents enabled partially by the effectiveness of the ranking technology. Few studies, however, attempt to investigate the impact of system components or mechanisms on user judgement of the search results. The closest in evaluative studies of search engines which investigate system features as an explanation for the results obtained and impact on user satisfaction are those which look at ranking.

3.2.1.1 Ranking

Courtois and Berry (1999) expressed their surprise in finding that little research has been done on search engines ranking of documents in response to simple search queries, given that "results ranking has a major impact on users' satisfaction with web search engines and their success in retrieving relevant documents." They go on to point out that whilst judging relevance of the first 10 to 20 retrieved items may be effective in determining precision, it is not how users use the result list. Rather they are more likely to scan the list and retrieve only selected documents. In their research they judged the ranking of search engines based on the criteria "all terms" (are documents that contain all search terms ranked higher?), "proximity" (are documents which contain all terms ranked higher where the search terms are contained as a contiguous phrase?), and "location" (are documents which contain all search terms ranked higher where the search terms are contained in titles, headings or metatags?). They then speculated on the linking of the results to search features of the systems. For example, Lycos performed well on the "all terms" criterion and the default use of the operator AND may have enabled this. AltaVista performed well on the "proximity" criterion which may be a result of its weighting for proximity in the ranking algorithm. The results for "location" were however reported to be low across all the search engines. Finally, they report that results varied widely by search topic in that some yielded consistent ranking while others produced lists with a few documents that contained all terms scattered among many that did not.

3.2.2 Usability

Usability studies aim to involve the user more in the evaluation in indicating the factors which influence IR interaction and provide some understanding as to why or how these impact on performance. Harter and Hert (1997, p.42) draw on the HCI literature for its definition as a measure of "system ability to provide an effective, efficient, satisfying performance of the users task". The usability of retrieval systems have been researched by a range of measures such as accuracy, error rate, action/ process variables (number of commands, descriptors, screens accessed, search cycles), retrieval (e.g, recall), user perceptions of ease of use and satisfaction. Further investigations have analysed the relationships of such measures with user characteristics such as cognitive abilities.

A range of system features may help a user to formulate a query and work with a system to obtain the desired results, possibly to attain the performance levels of recall and precision. A menu of options or a template in addition to the query box might offer assistance to users who are unfamiliar with creating effective search syntax. The relatively intuitive interfaces of some engines take into account that, on average, most people do not search effectively. Thus the intention is to prevent disappointment, or worse satisfaction with results retrieved from an inept search. Indeed, the interface (and non-retrieval devices) may affect the whole mode of interaction for the user and hence influence the demands the user indirectly puts on the back end search technology. Several listings and comparisons of search engine features, such as query formulation tools, can be found in the literature (such as, Dong and Su (1997); Feldman (1998); Kimmel (1996); Winship (1995)). These comparative listings do not however evaluate the features with respect to their impact on search performance, at least in any systematic or controlled manner. Such an evaluation of the functionality of interactive mechanisms would be desirable given the rapid advance of interface technology as a major area of research and development in these systems. The difficulty, however, is how (if performance is measured by some effectiveness measure(s)) to determine the impact of the back-end index and search mechanisms from the front-end tools which affect users' interaction and thus the demands that they make on the technology. All features of a system (and arguably contextual factors such as user characteristics) will have some impact on user interaction, searcher performance, and in turn on the actual system performance.

The problem posed for the evaluation of interactive systems is illustrated in the wide range of issues and interactive features studied under Interactive TREC. The Interactive track, added at TREC's 3rd annual conference, has the goal to develop evaluation methodologies for the interactive task (Harman 1996); that is, an investigation of the process as well as the outcome in interactive searching. Participants are encouraged to investigate different (user) approaches to conducting a TREC search task and investigate reasons for the results obtained. Researchers have used this venue to investigate a range of issues in comparing searcher performance using different systems/ interfaces. For example, investigations have been carried out on the use and utility of relevance feedback and ranking in interactive IR; the effects of topic order, difficulty, and domain on performance; the effects of using visualisation techniques; the extent to which searchers develop new searching behaviours; and, to investigate the effectiveness of different styles of interaction. *(*Voorhees and Garofolo, 2000)

In the context of TREC 8 interactive track, Fowkes and Beaulieu (2000) examined searching behaviour with a relevance feedback system to test a hypothesis that feedback would lead to better performance and searchers would prefer the system with relevance feedback. The findings on searching behaviour were related to the query formulation and reformulation stages of an interactive search process. Overall the norm was to use between 2 and 4 single query terms extracted from the given topic descriptions, and the queries were reformulated in only 15% of the searches. No statistical difference was found in the performance with retrieval with/without relevance feedback, and 75% of searchers did not perceive any difference between the two systems. Further analysis identified 3 levels of task-characteristics according to the degree of [searcher] interpretation needed to define a topic. This provided some understanding for how different task characteristics influenced search behaviour. Relevance Feedback came into play in different ways dependent on topic complexity. Automatic query expansion was found to be effective in improving simple queries but for more complex queries interactive query expansion with contributions from both searcher and system appeared to be more effective.

3.2.3 Searcher contexts

The realities of retrieval situations, as represented by the activities of users, define many contextual characterisations of users and tasks as parameters to be captured for an evaluation of interactive systems and facilities. Investigations which have sought to identify factors or (searcher traits as) predictors of search performance help to define these external variables of retrieval setups. With the shift from trained intermediaries to novice end users of IR systems came much research into the impact of individual differences on searcher behaviour and performance. For example, Saracevic et al. (1988) lists such research which study, for example, differences in users' search experience, training, cognitive characteristics, and perception of the information need on online searching. In their investigation of the nature of information seeking behaviour, Saracevic et al. examined five aspects of users, questions, searchers, searches, and outputs. A major outcome was the correlation of system performance with these variables thus identifying the external (user) factors that impact on search performance and which system designers should be aware. It is of interest that Beaulieu et al. (2000) analysed a further two stages of the interactive search process, *browsing and selecting documents, and viewing full documents*, with respect to user behaviour. Again it was found that different contexts influenced search behaviour. For example, two styles of browsing emerged in users' examination of the documents retrieved. When multiple answers to the query were found searchers worked systematically; but when few answers could be identified searchers were more selective. Scanning was seen to be the most prevalent strategy for evaluating document content, leading the researchers to posit that searchers seek considerable contextual information before making a relevance judgement. Further they state that passage retrieval, where the searcher is taken to the best passage that represents the highest scoring document section in relation to query term occurrence, was found to be disorienting and counter intuitive when searching on less familiar topics.

3.2.4 User satisfaction

Evaluation of a retrieval system's performance can thus be conducted in the abstract context of a Cranfield test or in an operational environment involving end users. The latter, as the above review indicates, presents serious challenges with the additional layers of complexity with respect to its design. The users' cognitive state (especially their understanding of the information need) will constantly change as they interact with the system and view documents. Such learning effects necessitate large and costly samples to replicate the search for system comparison. Furthermore the intrinsic variation in user needs and cognitive characteristics of the searcher are linked in some way to the relevance decisions they make, and to the use and value of different search facilities. A judgement of utility will be subjective and depending on what is important to the user different system features will impact on this judgement. Evaluation of interactive features (impact on search performance) must be undertaken in a complex test environment where searcher behaviour will impact on search performance and the context of the user's information need will affect the usefulness of the system search features. This all makes for generalisations and reliable comparisons about IR performance difficult.

An alternative approach to evaluation from a user perspective is to attempt to understand how users of the system themselves evaluate performance. The construct of user satisfaction used in system evaluation aims to achieve such a summary expression of users' perceptions based on the usefulness of a system. Throughout the history of evaluation, subjective measures concerning user satisfaction with search experience have been gathered. **Lancaster and Warner (1993)** report that such studies have consistently shown accessibility and ease of use to be the prime factors influencing the choice of an information source. Our review of user satisfaction and search engines (3.3) would seem to confirm the emergence of key influencing factors, but also reveals the multiple dimensions on which evaluation from a user perspective can be based. The reason for this complexity stems from the variety of criteria, based on user requirements, on which users may judge system success and the variations in user contexts which impact on an expression of system satisfaction. Tessier et al (1977) put forward three assumptions which they claim imply how satisfaction should be measured:

1) that user's satisfaction will be a function of how well the product fits their requirement;
2) that the user's state of satisfaction is experienced within the framework of their expectations;
3) that people may seek a solution within an acceptable range instead of an ideal or perfect solution.

The remainder of this review then seeks evidence for these assumptions which form the maxims for the proposed framework set out in Chapter 4 for the evaluation of search engines from a user perspective.

3.3 User satisfaction based evaluations of search engines

Stobart and Kerridge *(* 1996) revealed users' choice of engines to be dictated by *speed of access*, and other factors such as *size, habit, accuracy of data, user friendliness* and the *interface.* Nahl (1998) involved users in rating their self-confidence, stress level, understanding of the topic, satisfaction, and usefulness. It was found that *ease of use* and *fast response time* were important elements in determining self-confidence, stress and satisfaction levels. Furthermore Nahl concluded that "a search engine is perceived in the context of the information content it gives access to". This would seem to indicate that a user's perception of ease of use and thus value of a search tool is influenced by the extent to which the search results are of interest to the searcher. Nasois et al (1998) reported that the search results in their investigation of search engine capabilities were judged according to whether they would satisfy an easily pleased user or hard to please user. The suggestion is that user-traits may impact on the judgement of system success. Golovchinsky (1996) reported that users' view of recall increased with [increasing] number of articles displayed on the screen simultaneously. This would suggest that system characteristics may impact on users' perception of performance.

Su and Chen (1999) proposed a methodology for a dimensional approach to the evaluation of search engine performance from a user perspective. Based on a tested methodology (Su 1992, 1998) fifteen performance measures were grouped under five criteria of, relevance, efficiency, utility, user satisfaction, and connectivity. In recognition of the contributory factors of user characteristics in IR performance and evaluation, these were grouped under personal and educational backgrounds, user information needs/search requirements, and search strategies. Eleven participants were recruited to search for their topic on each of the four engines, AltaVista, Infoseek, Lycos and Opentext. Each participant made relevance judgements of the retrieved items and chose the five most relevant from the "top 20" and ranked them in decreasing order of relevance so that user and engine ranking of retrieved items could be compared. Participants were also interviewed to obtain ratings and reasons for satisfaction and utility. The pilot study found a number of differences among the 4 engines with none dominating in every aspect of the multi-dimensional evaluation. Lycos retrieved the highest number relevant and partially relevant items and had the highest mean precision ratio. However, in-spite of this, users assigned higher satisfaction ratings on precision for AltaVista. Lycos had the best rank correlation with users' relevance ranking. Although, AltaVista had the highest validity of links, satisfaction with online document, search interface and output format, and the highest value of the search results as a whole.

Wang et al (1999) approached the evaluation dimensions from a different viewpoint, that of the customer utilising the service. Their study was carried out using modified SERVQUAL dimensions to measure users' expectations and perceptions of search engines where good service quality is that which matches or exceeds expectation. Summarised here, these again show a number of dimensions and associated measurement criteria on which users might evaluate a search service. **Tangibles** (info is well organised; different search methods are available; a large amount of information is available; can narrow search topic). **Reliability** (good syntax consistency for the keywords in searching; search results are relevant to query). **Responsiveness** (search results are provided quickly). **Assurance** (no repartition of pages/sites; no dead links; information is up to date). **Empathy** (the layout on first impression is easy to understand; offers natural language searching; there are help screens, introductory pages or sample searches to guide the user; offers language selection for documents written in specific language). Preliminary analysis of service quality indicated that user understanding of and satisfaction with the quality of search engines are low. Among other suggestions they identify the biggest problem faced by searchers is the "needle in t he haystack" phenomenon and state "the ability to refine a query in a sensible way is very important to improving the quality of search engines" (p506)

3.4 Chapter summary

The evaluation of a search engine's performance in a controlled environment meets an important objective of the system, to retrieve relevant items for a given query. Its limitations, however, have focussed the question of how to evaluate search engines from a user-perspective based on the utility of the retrieved items and the usability of the system itself given the complex interaction of many user and system variables on performance. The use of user satisfaction as a surrogate measure has a long-standing tradition in evaluation studies. Based, however, on Tessier's assumption user satisfaction is a function of how well the system fits a user requirement, it follows that a variety of criteria may be used on which to base a measure of user satisfaction. Furthermore any measure of user satisfaction in itself is limited if there is no consideration of the system itself which has lead to a user's judgement only an assumption that users who have higher scores are using the better systems. In the majority of the studies reported above which used in combination objective measures of system performance and subjective measures of satisfaction seemingly contradictory results were obtained. For example, in Su (1998) it was reported that users with a low expectation of finding information expressed high satisfaction with a set of low precision results. It is proposed in our study that a framework for evaluation is needed if we are to make sense of such results which seem to confirm Tessier's assumption that a user's state of satisfaction is expressed in a framework of expectation.

Our conceptual framework for an evaluation of user satisfaction views a retrieval system as a means by which some individual performs some goal-directed task. To this end, user satisfaction is a multidimensional construct, which will vary across user and query contexts. The need to develop a framework for the evaluation of system contribution to the search process is articulated well in Belkin et al. (in Harter and Hert, 1997. p26) "*if we are going to serious about evaluating effectiveness of interactive IR, we need to develop … new performance measures. … that we develop measures based upon the search process itself and upon the task which has lead the searchers to engage in the IR situation*." User studies have begun to give some sense of what users are doing during IR interaction and provide models of valuable conceptualisations of the IR process. The reality however is that there is little consensus on what epitomizes the Information seeking phenomena, and by extension different perspectives on a model ma y lead to different focus for evaluation. For this reason, we draw on a general model of the information task and define user satisfaction measures within this theoretical view to focus in the evaluation on the degree to which system characteristics supports the user in their task needs. Our model suggests that users will give higher evaluations based not only on inherent characteristics of a system, but also on the extent to which that system meets their task needs and their individual abilities. Therefore a single system could get very different evaluations from users with different tasks, needs and abilities. Thus our framework for evaluation will incorporate a means by which we can evaluate the usefulness of a system with respect to the task the end user is undertaking.

Chapter 4. Development of the Framework

4.1 Introduction

Our aim is to develop a framework for the evaluation of Internet search engines from a user perspective. Towards this end we posit that user satisfaction is a complex multidimensional construct. It is not, however expressed by the user in the abstract but rather it constitutes some judgement of how well the service or technology fits a user's requirements. Thus in the framework for evaluation, user satisfaction must be defined within this theoretical basis to link system characteristics to their possible impact on the user task. In this section we present a preliminary construction of such a framework intended to structure existing measures and variables to provide a meaningful system evaluation from a user perspective. The small scale implementation described is not intended as an evaluation of the search engines used, as such, but rather as a means to test the feasibility of the proposed framework and to gather user data which may be used in its refinement towards an evaluation tool.

The framework proposed is conceptualised as *user satisfaction with a system is a function system-task fit and is expressed in a moderating context of user requirement*.

- *User evaluation criteria*

A general statement of the information retrieval task is that a user interacts with a retrieval system in order to retrieve specific items that will satisfy an information need. Based on a general model of the retrieval process we derive statements of user requirements, what a goal directed user is trying to do. Our premise is that meaningful user satisfaction measures can be obtained for system evaluation when defined within these dimensions of the IR task. That is, user satisfaction is an elicited response to the extent to which the system supports a task and can be evaluated by criteria which are related to what the user is trying to do suggested by the dimensions.

- *Measures*

To obtain some user evaluation along these dimensions, each criterion was operationalised by a set of measures which we considered reflected the user task. This perspective on user satisfaction measures enabled us to link in the system features which support the user in the retrieval task.

- *Context*

The framework proposed further seeks to incorporate a moderating context which may cause users to make different demands on the system which, it follows, will lead to varying user evaluation of the usefulness or functionality of the system features.

In the empirical investigation conducted with a view to developing such an evaluation framework we thus set out to better understand user evaluations of system satisfaction, that is how users are satisfied and on what criteria. User data were collected and analysed accordingly as follows:

⇨ User success ratings assigned on the four criteria, assigned by the task dimensions, were correlated with an overall success judgement to find which, if any, appears be the most important factor in defining user satisfaction.

⇨ Users ratings on the measures used were correlated with the overall success ratings for each associated criterion to find which measure, if any, contributed most strongly to the user's overall rating of a dimension.

⇨ User derived reasons for attributing satisfaction ratings, overall and on each criterion, were collected using open-ended questions and analysed to validate, or otherwise, our measures as those which users themselves base an evaluation of system satisfaction.

Whilst the test did not set out to evaluate the impact of system features on users' ratings, in the spirit of a feasibility investigation we did seek to find whether users' expression of satisfaction were simply random or whether they were meaningful evaluations of the given system characteristics. By basing our measures on system features which may support a user in a task dimension we expected to observe variations of satisfaction ratings across search engines which differ in the way they support the retrieval task.

⇨ 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.

In the framework proposed it is suggested that user evaluation of the system may be moderated by some contextual characterisation of user and information query. The impact of this context on user satisfaction was explored in the testing of the framework. That is, we wanted to find if our characterisation of context led to systems receiving different evaluations. The contexts that seem to have the most impact could be used to develop the framework for a system evaluation which links system features with user evaluations as dependent on certain user/task contexts or under which context a system obtains higher rating, and thus features supporting certain tasks.

.

⇨ The four task identifiers (task defined, task purpose, task knowledge, and task probability) were analysed against the overall satisfaction ratings and the four evaluation criteria across all four search engines to ascertain if a moderating effect of context was obtained.

This chapter sets out the development of the proposed framework for evaluation and details its implementation for the feasibility study. Since the investigation is not intended to be an evaluation of the search engines as such we refer to the engines used in the study as SystemA, SystemB, and SystemC.

4.1.1 IR task/process models

The identification of a process model, which proposes assumptions as to what the user is trying/wants to do, provides the rationale for our measures. The specific task domain is, *users wish to retrieve relevant items to satisfy their information need*. Although individuals' information seeking goals can differ quite widely, standard models of the information seeking process contain the core steps of query specification, receipt of results in an interactive cycle. The process model on which we draw (Salton, 1989) identifies interacting steps, which are not necessarily sequential and may be repeated. This gives the dimensions on which users might evaluate system success/ satisfaction. These are

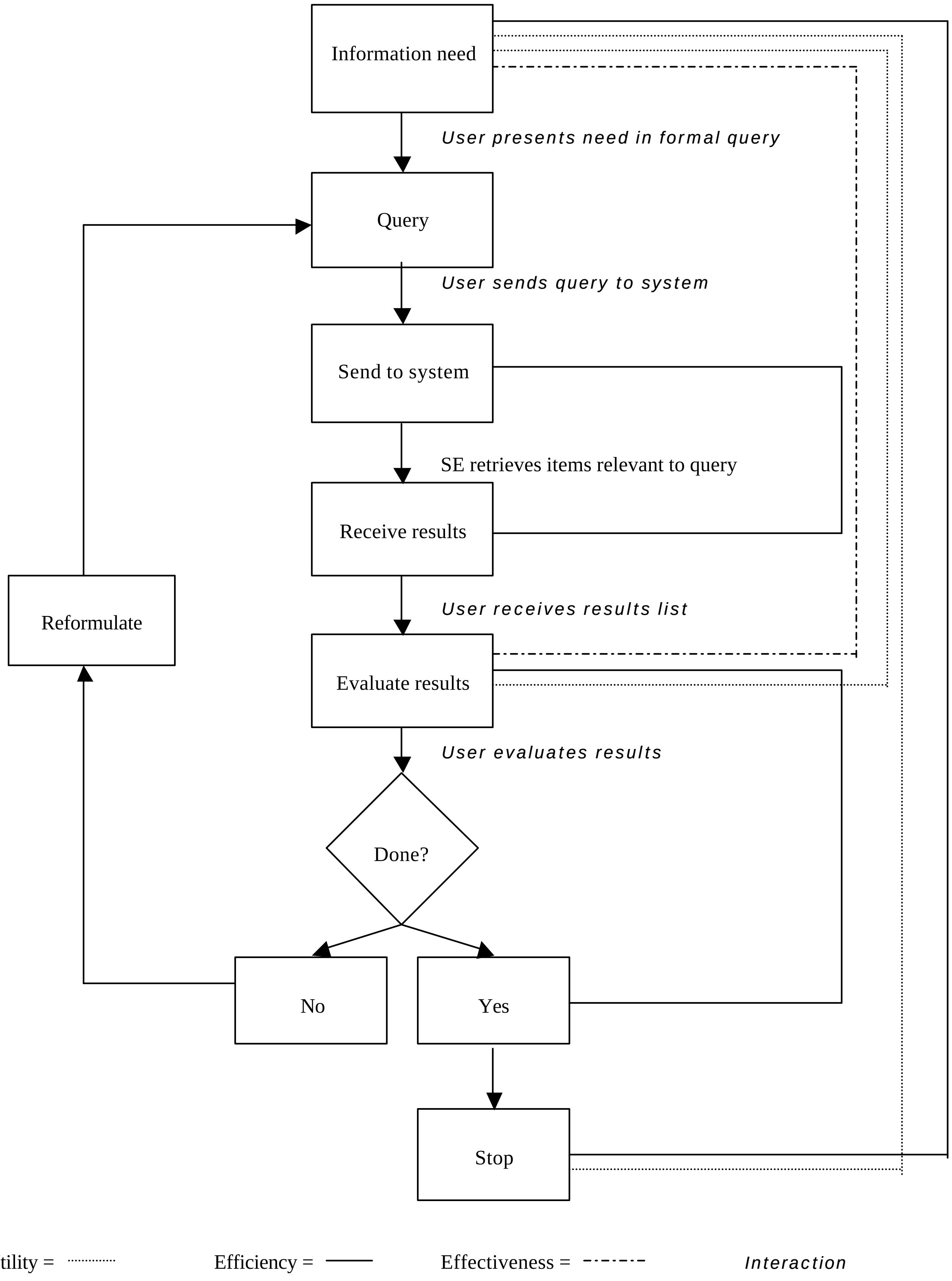
1. *Users will formulate/submit a query;*
2. *Users will receive results;*
3. *Users will evaluate results – end or modify (Note, a possible feedback loop here); and*
4. *Users will evaluate success of the search as a whole*

For each dimension, we can relate the criteria, *Effectiveness, Efficiency, Utility* and *Interaction*, by which users might evaluate system satisfaction on these task dimensions.

Dimension 1 *Users will formulate/submit a query* evaluated on the criterion of **interaction (query)**
Dimension 2 *Users will receive results* evaluated on the criterion of **interaction (output)**
Dimension 3 *Users will evaluate results* evaluated on criteria of **effectiveness & relevance/ranking**
Dimension 4 *Users will evaluate success of the search* evaluated on criteria of **efficiency & utility**

We justify the use of this standard model in that it describes the basics of the retrieval process. However, it is important to note that this model has been contrasted with others, such as Bates' (1989) berrypicking model which challenges the view that the information need will remain static throughout the process and that the main value of the search resides in a set of retrieved documents. This alternative model then emphasises the interaction which takes place whereby a user learns, goals are triggered, and information acquired along the way.

The task model, based on a simplified model of the information access process (Hearst, 1999):



4.1.2 Measures

Measures were developed which defined user evaluations for each criterion along each of the dimensions. Each criterion was thus unpacked to the group of variables on which user satisfaction with an interactive retrieval system can be measured. The measures were identified in the process of defining each criterion when related to the IR task/process dimensions. The intention being to develop user satisfaction evaluation variables which, in the framework, relate to system function (features) and will provide for a system rating based on task fit in an end user searching environment. The majority came from existing (and generally accepted) measures. Those developed for the proposed evaluation framework were mapped, in a sense, to the (function of) system features which supported the dimension in question.

Table 8 Framework for the evaluation of SEs from a user's perspective

Dim1 User satisfaction with **Effectiveness** (SE features)	Dim2 User satisfaction with **Efficiency** (SE features)	Dim3 User satisfaction with **Utility** (Output)	Dim4 User satisfaction **Interaction** (Interface)
1.1 Precision1 (traditional measurement) 1.2 Precision2 – user satisfaction with precision 1.3 P3, comparison of P1 and P2 1.4 Ranking1 – system 1.5 Ranking2 – user satisfaction with ranking 1.6 R3, comparison of R1 and R2	2.1 Search session time 2.2 Response time 2.3 Relevance assessment time (in situ)	3.1 Value of search results as a whole 3.1.1 Satisfaction with results 3.1.2 Resolution of the problem 3.1.3 Rate value of participation 3.1.4 Quality of sources 3.2 Validity of links 3.3 Number of links followed up	4.1 User satisfaction with output display / visualisation of representation of item 4.1.1 User satisfaction with manipulation of output 4.1.2 User satisfaction with visualisation of representation of item 4.2 User satisfaction with interface 4.2.1 User satisfaction with query input 4.2.2 User satisfaction with query modification 4.2.3 User satisfaction with query visualisation/clarification

4.2 Task Dimensions and User Measures

Four dimensions were identified from IR task/process models, and were used as the basis on which to suggest the criteria on which users might evaluate or rate system

Dimension1 *Users will evaluate results*

Criterion *Retrieval performance (effectiveness) will affect user evaluation of SE*

Measures of retrieval effectiveness are based on the notion of relevance, and are based on the assumption that given a document collection and a query some documents are relevant to the query and some are not. The objective of the IR system is to retrieve relevant documents and to suppress the retrieval of non-relevant documents. System output can then be evaluated on the basis of how well these objectives are met. Most used are the measures of recall and precision. A user's evaluation of a SE will be partially dependent on the ability of the system to meet these basic criteria.

In a web based environment with direct user interaction these traditional measures may not be appropriate, instead other relevance based measures may be used to provide criterion for evaluating effectiveness in the performance of the system. That is, relevance will not be measured on binary (relevant/non-relevant) scale but instead the concept of relevance will encompass non binary judgements relative/partial differentiated into situation 'usefulness' or 'utility or topicality' – that is assessment categories viewed as dimensions of information needs.

Measures *This dimension can be evaluated by*:

1.1 Precision1 (traditional measurement)
1.2 Precision2 – user satisfaction with precision
1.3 P3, comparison of P1 and P2
1.4 Ranking1 – system
1.5 Ranking2 – user satisfaction with ranking
1.6 R3, comparison of R1 and R2

In the empirical investigation data was gathered and analysed on the measure based on Precision2 – user satisfaction with precision and Ranking2 – user satisfaction with ranking. 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 each participant assesses. Participants were then asked to rate on a five point scale their satisfaction with the precision of the search results. Satisfaction with ranking order was obtained on five point scale. 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.

Dimension2 *Users will evaluate success of the search as a whole*

Criterion *Efficiency will affect user evaluation of SE*

Efficiency seems a little hard to define, but basically is concerned with how efficient the system is in retrieving the required information. Boyce et al (1994) 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.” (p.241). They also state that if the service or good is not judged to be effective then efficiency has little meaning. Also, Dong and Su (1997) state that “response time is becoming a very important issue for many users” (p.79). Therefore, a user’s evaluation of a SE will be affected by the system’s efficiency. The premise being that users want to retrieve information as efficiently as possible, which may in part equate to as quickly as possible.

Measures *A user will evaluate the dimension by:*

2.1 Search session time

2.2 Response time

2.3 Relevance assessment time (in situ

In the empirical investigation the search session time was noted and used in the analysis. Participants were asked to rate on a five point scale the overall success of the search engine in retrieving items efficiently.

Dimension3 *(Output)*

Criterion *Utility will affect user evaluation of SE*

Authors have also argued for other measures, such as utility so that an information system is evaluated on the basis of how useful it is to its users. Utility has been defined as “the degree of actual usefulness of answers to an information seeker” (Saracevic and Kantor, 1988, p.169). Utility measures are based on users’ expressions of degree of satisfaction and value of the retrieved items as a whole. The utility approach highlights many factors, other than relevance, which will affect a user’s evaluation of the system’s performance. Cleverdon (1991) argued (with Cooper, 1973), who put forward a straight utility-theory single measure) that retrieval effectiveness measures should be used in combination with more user-oriented measures. The aim being to produce an evaluation on utility with other such factors as subjective satisfaction statements, search costs, time spent etc which could be related to nominal recall and precision values in such a way as to indicate how various parameters ought to be changed

It is clear that utility is concerned with the degree of actual usefulness of retrieved items to the user, yet Saracevic and Kantor (and Su, 1998, p.558) report that standard utility measures do not exist. Saracevic and

Kantor 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.* Indeed, various factors may bear on users' judgements of overall satisfaction with the value of the search results. For example, depending on the user and their information need, 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 specific, or at the right level, to user need. In the context of web evaluation we supplemented an 'overall satisfaction with value of search results' with the three variables of Validity of links, Number of links followed up, and Quality of sources which may impact on user evaluation of satisfaction with the search engine [3]

Measures *A user will evaluate the dimension by*:

3.1 Value of search results as a whole

3.1.1 Satisfaction with results

3.1.2 Resolution of the problem

3.1.3 Rate value of participation

3.1.4 Quality of sources

3.2 Validity of links

3.3 Number of links followed up

In the empirical investigation participants were asked to rate on a five point scale the worth of their participation, with respect to the information which resulted; 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.

Dimension4 *Users will formulate/submit a query*
Users will receive results

Criterion *Interaction will affect user evaluation of SE*

Interaction is a concept which is often discussed but little defined. In the context of web SEs it is how the user directly interacts and manipulates/commands the system to retrieve the information or specific items they require. Interaction will be largely determined by satisfaction measures alone. Belkin and Vickery (1985) state "satisfaction as a criterion for evaluation of information systems is a concept explicitly intended

[3] ESL calculates the cost paid by a user in the sense of the number of sites the user must look through before finding sufficient items to satisfy the query. This seems to incorporate some of the measures which we have under Efficiency and Utility.)

to extend the range of factors relevant to the evaluation. In particular, the intention is to move away from evaluation according to system performance, the basis of information retrieval, and toward an overall judgement based on user reaction to the system" (p.194).

Based on the features of search engines which might support a user in the IR task of submitting/ formulating a query, we defined the measure of 'user satisfaction with interface' as comprising measurement on three variables of user satisfaction with query input, query modification, and query visualisation. User satisfaction with query input may be influenced by the perceived ease by which the user can express a query. For example the availability of different search methods, such as natural language searching or power search to narrow a search topic. User satisfaction with query modification may be influenced by assistance provided in formulating the search, such as suggesting query terms or offering a feedback mechanism. User satisfaction with query visualisation may be influenced by any provision in helping the user in understanding the impact of a query. An obvious example is the use of folders which could have multiple impact on the user's understanding of the query, such as suggesting different perspectives of the topic or information which might be useful in a different search.

On receiving results the user will be involved in some process of interpreting the results in the given frame of the information need. On a general level users would want to easily see why an item was retrieved and to quickly see its meaning to make a relevancy judgement. Features of a search engine which might support a user in this task lie in its summary representation of items for visualising the 'aboutness' of item, and extent to which information is presented in clear and organised manner . We defined the measure of 'User satisfaction with output display' as comprising measurement on the variables relating to manipulation of the output (e.g. summary display features (category labels), sort by) and visualisation of item representation.

Measures *A user will evaluate the dimension by:*

4.1 User satisfaction with output display
4.1.1 User satisfaction with user manipulation of output
4.1.2 User satisfaction with visualisation of representation of item

4.2 User satisfaction with interface
4.2.1 User satisfaction with query input
4.2.2 User satisfaction with query modification
4.2.3 User satisfaction with query visualisation/clarification

In the empirical investigation data was gathered on user satisfaction on all of these measures using a five point scale.

4.2.1 Context

There are many factors which could be used to characterise the user context by, for example, user traits, experience, background, cognition; the information request, subject, type users expectation, perception or understanding of the request. (*Note, also searcher behaviour , search strategies, tactics will make different demands affect performance and thus user evaluation)* Sitting on top of our model of the IR task process used in the development of the evaluation criteria is the context that users have an information need. Thus in our evaluation framework as possible moderation of user evaluations we characterise this context by factors such user intent/ amount of prior knowledge/ expectation .

User context was characterised by responding to the questions

- Problem definition scale *(in your opinion, and on a scale from 1 -5, would you describe your problem as weakly defined or clearly defined?)*
- Intent scale *(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?)*
- Internal knowledge scale *(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?)*
- Problem-public knowledge *scale (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?)* (Saracevic and Kantor, 1988)

In addition, we incorporated such questions as (from Koll, 2000): Searching for known item; Searching for an unknown item; Searching for any item; Searching for the most relevant item; Searching for most of the items; Searching for all of the items; Searching for affirmation that there are no items; Searching for like items; Searching for new items to supplement items already obtained previously.

4.3 Implementation

Twenty three participants were recruited from second year students of the Department of Information and Communications, MMU. A short introduction was given to the participants a few days prior to their search session to explain to them the project to which they were contributing and to present them with the Information Need and User Characteristic questionnaire which they were required to complete before their search session. No restrictions were placed on the type of information they required or the purpose for which it was intended. The following table presents participants' characteristics.

Table 9 User characteristics

Characteristic	Variable options	Number of cases	Percentage
Gender	Male	7	30
	Female	16	70
Age	18-20	6	26
	21-30	6	26
	31-40	7	30
	41-50	3	13
	51-60	1	4
	61-70	0	0
	70+	0	0
Academic status	2nd year	23	100
IR experience	None	1	4
	Some	16	70
	Lots	6	26
Computer experience	None	0	0
	Some	14	61
	Lots	9	39
Internet experience	None	0	0
	Some	14	61
	Lots	9	39

The participants were split into two groups searching on two different days. On arrival each student was given a second questionnaire (three copies – one for each SE) which was concemed with participants' ratings of dimensions and measures of each SE. They were instructed to read this before commencing searching. Each student was required to search three particular SEs in an order specified by the Test Administrator, the order of which was varied to remove learning curve effect by a 3x3 Latin square. Therefore, each SE was searched in each of the three positions by an equal number of participants. A short introduction was given to the participants prior to searching. The introduction included: 1) order of SEs to use; 2) how to print-out results; 3) how long to search for (free choice) and, 4) what to search for (free choice).

Participants were asked to search for an information need of their choice, to use as many reformulations as required and to search for as long as they would under normal conditions. This was to be repeated on the remaining two SEs. Once they retrieved a set of results, i.e. a hitlist, they were asked to print these out. From this hitlist they made relevance judgements which they marked on the printout and handed these in with their completed questionnaires. These relevance judgements were based on a set of guidelines given to each participant before searching. These guidelines defined relevance in terms of a three point scale where R = relevant, PR = partially relevant and NR = not relevant.

In some instances participants were unable to complete three whole searches within the time of the session (two hours). In these cases the Test Administrator accepted only a completed test on a SE. Fifteen participants completed the test on all three SEs, one participant completed the test on two SEs and seven participants completed the test on one SE. In this way 54 searches were collected during the test.

4.4 Data Analysis

Proposition 1

Our primary aim was to test the assertion that users' evaluation of a system based on satisfaction measures is multidimensional, that is overall satisfaction is not a single construct but a response to how well the system has supported the IR task which may be made on many dimensions.

In the quest to better understand how users evaluate these systems we asked users to give a system an overall success rating. By correlating ratings assigned on the four criteria, as suggested by the task dimensions, we aim to find which, if any, appears be the most important or contributory factor to users' overall judgement.

Table 10 Global and SE level - Overall success rating correlated against the four criteria

Criterion	Correlation coefficient			
	Global	SystemA	SystemB	SystemC
Effectiveness	.759**	.779**	.795**	.729**
Efficiency	.817**	.843**	.908**	.741**
Utility	.710**	.362	.930**	.806**
Interaction	.592*	.511*	.660*	.580*

* = moderate strength correlation

** = strong correlation

Globally he criterion with the strongest correlation with users' overall rating is Efficiency, followed by Effectiveness, Utility and Interaction. On SystemA the strongest correlation is Efficiency, followed by Effectiveness, Interaction and Utility – where a weaker correlation is demonstrated. On SystemB the strongest correlation is Utility, followed by Efficiency, Effectiveness and Interaction. On SystemC the strongest correlation is Utility, followed by Efficiency, Effectiveness and Interaction.

The strength of the correlation ratings assigned on the four criteria with users' overall success judgement in this study indicates that user satisfaction is a multidimensional construct and that the measures used were valid. That is, a user judgement of system satisfaction is based on a response to the extent to which the system supports the many dimensions of the IR task. The Efficiency criterion held the strongest correlation with the success judgement, and the Interaction criterion held the lowest. This could suggest that efficiency is the most important criterion in the users' minds when assigning a success rating, and that the users in this study have little interest in system interaction.

By further correlation of measures within each criterion we ask does a user's (low or high) rating on a single variable lead to a low/high rating on the related criterion

Table 11 Global and SE level - Measures correlated within Effectiveness criterion

Measure	Correlation Coefficient			
	Global	SystemA	SystemB	SystemC
Satisfaction with relevance	.733**	.864**	.639*	.794**
Satisfaction with ranking	.485*	.620*	.371	.541*

* = moderate strength correlation ** = strong correlation

Both Globally and at SE level the strongest correlation between Effectiveness and individual measures is satisfaction with relevance, followed by satisfaction with ranking.

Table 12 Global and SE level - Measures correlated within Efficiency criterion

Measure	Correlation Coefficient			
	Global	SystemA	SystemB	SystemC
Time taken in minutes	.062	.018	-.150	.336

* = moderate strength correlation ** = strong correlation

From these results it can be seen that a negligible correlation between time taken to search and the Efficiency criterion exists.

Table 13 Global and SE level - Measures correlated within Utility criterion

Measure	Correlation coefficient			
	Global	SystemA	SystemB	SystemC
Rate value of participation	-.557*	-.600*	-.394	-.682*
Resolution of problem	.723**	.756**	.782**	.687*
Satisfaction with results	.755**	.552*	.963**	.769**
Value of results as a whole	.742**	.594*	.913**	.772**
Overall quality of results	.702**	.512*	.804*	.833**

* = moderate strength correlation ** = strong correlation

Globally the strongest correlation between an individual measure and Utility is *satisfaction with results*. The *rate value of participation* measure has a negative correlation which indicates that as Utility rises the value of participation decreases. On SystemA the strongest correlation is *resolution of the problem*; on SystemB the strongest correlation is *satisfaction with results;* and on SystemC the strongest correlation is *overall quality of results*.

Table 14 Global and SE level - Measures correlated with Interaction criterion

Measure	Correlation coefficient			
	Global	SystemA	SystemB	SystemC
Importance of ability to change output	.132	-.186	.345	.171
Ease of understanding item/s from hitlist	.452*	.400*	.524*	.431*
Satisfaction with input facility	.486*	.361	.600*	.488*
Importance of ability to modify query	.437*	.476*	.656*	.305
Satisfaction with presentation of query	.506*	.573*	.427*	.524*
How helpful was Help	.190	.286	-.027	.325

* = moderate strength correlation ** = strong correlation

Globally the individual measure with the strongest correlation is *satisfaction with presentation of the query* which is of moderate strength. This is followed by satisfaction with facility to input query, ease of understanding item/s from the hitlist, importance of ability to modify query, how helpful was Help and importance of ability to change output. These latter two demonstrate weak correlations.

On SystemA the strongest correlation is satisfaction with *presentation of query*, while satisfaction with input facility, how helpful was Help and importance of ability to change output, demonstrate weak correlations. On SystemB the strongest correlation is importance of *ability to modify query*, while importance of ability to modify query and how helpful was Help show a weak correlation. On SystemC the measure with the strongest correlation is *presentation of query*, with how helpful was Help, importance of ability to modify query and importance of ability to change output demonstrating weak correlations.

The implementation of the test was intended to be exploratory of user evaluations and the framework rather than an evaluation of the systems as such. For this reason we sought to validate our measures as those which are important from a user perspective when evaluating or making some judgement of a system. Towards this end we included open-ended questions to collect user-derived reasons for attributing satisfaction rating with the system as a whole and rating for each dimension. In-depth analysis, such as categorisation of the some 250 comments, was considered to be beyond the scope of this feasibility study. These comments are, however, used to substantiate our interpretation of the above data analysis.

The correlations of the user ratings on the measures within each criterion would seem to indicate the validity of these measures. User satisfaction with relevance held the strongest correlation with user ratings of the criterion Effectiveness. User-derived reasons for assigning ratings of success on this criterion would seem to confirm this finding.

"Information retrieved was extremely relevant to my needs"
"Although not all items retrieved were relevant those that were were very important ones"
"Most items retrieved appear to have some relevance"

"All items were of a certain amount of relevance
"Too much irrelevant information"

The measure of search time, however, held a low correlation as a measure of Efficiency. That the Efficiency criterion held the strongest correlation with an overall success judgement suggests that users define system efficiency as something other than the time taken to obtain search results. The user-derived reasons for assigning ratings of success on this criterion would indeed seem to suggest that users relate efficiency to the amount of effort required from themselves to conduct a search. For the purpose of this feasibility study this finding has implications for the further development of user measures in the evaluation framework which will be discussed in the conclusions.

"Ease of use"
"Had to redefine search twice"
"The search terms were attempting to pin down a concept that was hard to verbalise/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 locate 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"

The Utility measures all held strong correlations both globally and across the search engines. Our user-derived reasons would also indicate that these were measures which users themselves used in judging system performance. Further analysis would be required to ascertain if in fact all these measures were simply variations of the same measure "satisfaction with results".

"Those items found were useful"
"I have gained further info on the subject I was search for"
"Current and up to date info was located"

The correlations found with the Interaction measures were relatively low with user satisfaction with query presentation holding the strongest correlation. The user derived reasons, however, indicated that there was perhaps some expectation from the users that the system would provide some assistance in modifying the query and that this would impact on their evaluation of system interaction.

"I changed the query once and it was helpful"
"The SE easily allowed the query to be modified"
"I didn't like the style of layout of retrieved item"
"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 SystemB SE I usually use"
"Could lead to different routes of enquiry from the initial search term"

Proposition 2 *Characteristics of information systems will affect user evaluation on task dimensions*

The view taken is that user evaluations are not random, but reflect the characteristics of the system in supporting the users' task. For the purpose of the feasibility study we looked to find evidence that variation across the systems is found in the users' evaluation based on the measures used.

The strongest correlation with users' overall rating and Utility was found on SystemB, with the measures of user satisfaction with results, value of results as a whole, and overall quality of results correlating most strongly with user judgement on this criterion. This would suggest that users' high rating of Utility lead to a high rating of system success. In contrast, a very weak correlation was found on SystemA with users' overall rating and Utility. The marked difference in the strength of correlation found between the systems is interesting but only in that it suggests that users overall judgement of system success may be more strongly associated with a judgement made on a particular task dimension/ criterion depending on the system. In an 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, in our feasibility study using a small sample it was found that 68% of the users rated the utility of the results from SystemC as contributing very little to the resolution of the problem, and 58% of the users expressed dissatisfaction with the results as a whole. This could be compared with the 29% who expressed dissatisfaction with the results from SystemB.

Following this line of analysis for the Interaction dimension we can note that the strongest correlation with users' overall rating and Interaction was found on SystemB, with the measures of user satisfaction with input facility and ability to modify query correlating most strongly with user judgement on this criterion. In the analysis of central tendency it was found that 77% of users rated the ability to modify query as important with SystemB.

Proposition 3 *Query characteristics will affect user evaluation of SE*

In the framework proposed it is suggested that user evaluation of the system may be moderated by some contextual characterisation of the user and information query. That is, a user context makes different demands on system and thus lead to higher or lower user evaluations of satisfaction with system on dimensions of the information retrieval task.

For the purpose of the feasibility study we sought confirmation or otherwise that the query context will have some moderating effect on the evaluations. To this end 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 (General Feelings) and the four criteria. Again we stress that our study was not an evaluation of the systems but rather a testing of the feasibility of the framework as an evaluation tool. 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 between *task definition* and General Feelings (.407), Effectiveness (.418) and Efficiency (.482) were found indicating that as task definition increases so does overall satisfaction, satisfaction with Effectiveness and satisfaction with Efficiency. The correlations between *task definition* and Utility (.307) and Interaction (.221) were weak. Weak/very weak correlations were obtained between *task purpose*, *task knowledge*, and *task probability* and the overall satisfaction 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 and will work effectively with the system to obtain the results. The effect of the moderating context will be of more interest if notable variations can be found between the systems evaluated.

Again in the limitations of the feasibility study it is noted that the strongest correlations of *task defined* against Effectiveness (.561) and Efficiency (.702) were found on SystemB. Whilst weak correlations were found globally for *task purpose*, when based on the data obtained for SystemB moderate strength correlations were found for *task purpose* with Efficiency (.636) and Utility (.577). The comparison, for example, that on SystemA the correlations were weak, *task purpose* and Efficiency (.161) and Utility (.002), sets up a line of enquiry as to why a relatively strong correlation was obtained on SystemB. Again it would not be surprising if a correlation was found for 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. Further analysis of task purpose association with the Interaction measures revealed that correlations of moderate strength were obtained only with the measures ease of understanding item/s from hitlist and satisfaction with query presentation on SystemB. This may suggest that the features of SystemB which are related to the visual organisation and representation of the search results better support the user with a broad query.

4.5 Discussion and Conclusions

The aim of this project was to develop a framework for the evaluation of Internet Search Engines with an emphasis on a user-centered perspective. The review of search engine developments revealed a range of indexing and retrieval techniques which are employed to assist casual users in their task of retrieving information. In particular a range of novel search engine features or characteristics can be seen in the areas of search assistance (query formulation, modification, and visualisation), and results ranking. In this context, it is critical that we have some means to measure the impact system features have on users' satisfaction with respect to what they want to do or achieve with these systems. The review of approaches for the evaluation of retrieval systems served to highlight the complexity of evaluation studies which aim not only to obtain some objective measure of performance but also some measure of the utility of the retrieved results and the usability of the system from a user perspective. Consideration of this complex evaluation situation led us to the proposal of a conceptual framework for system evaluation in which user satisfaction is characterised as a function of system-task fit expressed in a moderating context of the user requirement. Thus the aim of this project was to explore the feasibility that the framework captures the complex interrelations among system and contextual parameters in such a way so as to provide meaningful user evaluation of the system. Towards this end we focussed our research on the definition of the construct of user satisfaction which in the proposed framework would be taken to be the dependent variable. The key to its use as a meaningful measure of the system and its functionality is the view that user satisfaction is a multidimensional construct, a function of the user's task requirement.

For the feasibility study a framework for evaluation was developed along these lines drawing, in the main, on existing user satisfaction measures and contextual characterisations. The general task model of the retrieval process provided the dimensions of user satisfaction and to an extent allowed us to identify the system components or features which may impact on the users' judgement of satisfaction with respect to task-support or fit. For the further development of user satisfaction measures use will be made of alternative models which give emphasis to the dynamic and interactive nature of the retrieval task. The main objective of the feasibility study was to test the notion that satisfaction is a multidimensional construct and the validity of the measures used.

The analysis of the data collected in the empirical investigation appears to support the notion that user satisfaction is expressed as a multidimensional construct with correlations held among the measures and overall success ratings. To ascertain the validity of the measures used, or to develop new user satisfaction evaluation statements, will require further analysis, in particular of the user-derived reasons for attributing system satisfaction on the dimensions of Efficiency and Interaction.

The conceptual argument underlying the evaluation framework is that user satisfaction, the strength of user evaluations, will be dependent on the system characteristics in supporting the associated task dimension, given the context of task demands and capabilities of the user. In the feasibility study some variation was found in the users' ratings on the criteria and measures across the search engines indicating that the features of the engines may have in some way contributed to users' evaluations of the systems. It is further possible to speculate that system characteristics such as 'selection of items for inclusion in database' may impact on the Utility judgement, and 'facilities for query modification, such as relevance feedback or suggesting terms' may impact on the Interaction judgement. Some impact of the identified query context was also found suggesting its moderating effect on users' evaluations of the systems. Again in the constraints of the feasibility study we could only speculate with great caution that a characteristic of the system may have better supported a particular query context. In a full-scale evaluation study appropriate statistical techniques, such as regression analysis, would be necessary to explore the relationships held among dependent and moderating variables and to express the overall performance measure as a function of these variables. The implementation of such an evaluation framework would require a far greater sample size than the one used for this feasibility study.

Our preliminary findings have revealed the complexity of the construct of user satisfaction as a measure of system performance, but also have indicated to us the potential value of the proposed framework for the evaluation of search engines. We therefore tentatively suggest that with further refinement the proposed framework will provide for a multidimensional user evaluation of search engines and may allow some evaluation of the specific features of search engines from a user perspective. Furthermore, the incorporation of a moderating context in the evaluation may provide a better understanding of the differences found in users' evaluations of the same system. Ultimately our aim would be to develop the evaluation framework so that not only is variation found across systems in users' expression of satisfaction but also that system characteristics can be identified which provide an explanation for variation within the evaluation dimensions and across the users' task contexts. Towards this end, research will continue at the Manchester Metropolitan University to develop a set of user evaluation statements which define the multidimensional construct of user satisfaction by the task dimensions of information retrieval.

References

AltaVista search: questions (1999) [Online] (URL: www.altavista.digital.com/av/content/ques_master.htm

Andrews, W. (1996) Search engines gain tools for sifting content on the fly. *Web Week*, 2(11), 41-42.

Back, J. and Summers, R. (unpublished results in Oppenheim, C., Morris, A., McKnight C., and Lowley S. (2000). The evaluation of www search engines. *Journal of Documentation*, 56(2), 190-211.

Barry, C.L., and Schamber, L. (1998) Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, 34(2/3), 219-236.

Bates, M. J. (1989) The design of browsing and berrypicking techniques for the on-line search interface. *Online Review*, 13(5), 407-431.

Belkin, N.J. and Vickery, A. (1985). *Interaction in information systems: A review of research from document retrieval to knowledge-based system* (p. 188-198). London: the British Library.

Berghel, H. (1997). Cyberspace 2000: Dealing with information overload. *Communications of the ACM*, 40(2), 19-24.

Boyce, B. et al (1994) *The measurement of information science*. Academic Press.

Chu, H. and Rosenthal, M. (1996) Search engines for the world wide web: A comparative study and evaluation methodology. *ASIS '96: Proceedings of the 59th ASIS annual meeting*, 33, p.127-135. Medford, NJ: Information Today. Available at http://www.asis.org/annual-96/ElectronicProceedings/chu.html

Clarke, S.J. and Willett, P. (1997). Estimating the recall performance of Web search engines. *Aslib Proceedings*, 49(7), 184-189.Andrews, W. (1996). Search engines gain tools for sifting content on the fly. *Web Week*, 2(11), 41-42.

Cleverdon, C.W. (1978). User evaluation of information retrieval systems. In: King, D.W (ed.). *Key papers in design and evaluation of retrieval systems*. New York: knowledge Industry.

Cleverdon, C.W. (1991). The significance of the Cranfield tests on indexing languages. *SIGIR '91. Proceedings of the ACM Special Interest Group on Information Retrieval*. 14th Annual International conference on research and development in information retrieval. Oct 13-16, 1991. 3-12.

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, 30-41.

Cooper, W.S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24, 87-100.

Courtois, M.P. and Berry, M.W. (1999) Results ranking in web search engines. *Online*, 23(3) Available at http://www.onlineinc.com/onlinemag/OLtocs/OLtocmay4.html

Croft, W.B. (1995). What do people want from Information Retrieval. *D-Lib Magazine*, November. [Online] (URL: www.dlib.org/dlib/november95/11croft.html)

Ding, W.and Marchionini, G. (1996). A comparative study of web search service performance. In: *ASIS 1996 Annual Conference Proceedings,* Baltimore, MD, Oct 19-24, 136-142.

Dong, X., and Su, L. (1997). A comparative study of web search service performance. *Proceedings of the American Society for Information Science*. 136-142.

Ellis, D. (1984). The effectiveness of information retrieval systems: the need for improved explanatory frameworks. *Social Science Information Studies*, 4, 261-272.

Ellis, D., Ford, N., and Furner J. (1998). In search of the unknown user: indexing and hypertext and the world wide web. *Journal of Documentation*, 54(1), 28-47.

Excite info: getting listed help (1999) [Online] (URL: http://www excite.com/Info/listing.html)

Feldman, S. (1998). Web search services in 1998: trends and challenges. *Searcher*, 6(6), 29-39. Available at: http:// www.infotoday.com/searcher/jun98/story2.htm

Feldman, S. (1999). Search Engines: the 1999 conference. *Information Today*, 16(6). Available at http://www.infotoday.com/it/jun/felman.htm

Fowkes, H. and Beaulieu, M. (2000). Interactive searching behaviour: Okapi experiment for Trec-8. *The BCS/ IRSG 22nd Annual Colloquium on Information Retrieval Research*, Cambridge, 5-7 April.

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

Gauch, S. and Wang, G. (1996) *Information fusion with ProFusion*. Webnet 96 Conference, San Francisco, CA, Oct 15-19. [Online] (URL: http://www.csbs.utsa.edu:80/info/webnet96/html/155.htm)

Golovchinsky, G. (1996) *From information retrieval to hypertext and back again: the role of interaction in the information exploration interface*. PhD Thesis. University of Toronto. http://anarch.ie.utoronto.ca/people/golovch/thesis/final/

Goodhue, Dale. L. (1995). Understanding user evaluations of information systems. *Management Science*, 41(12), 1827-1843

Harman, D. (1995). Overview of the second text retrieval conference (TREC-2). *Information Processing and Management*, 31, 271-289.

Harman, D. (1996). *The fourth text retrieval conference (TREC-4), NIST special publication*, p.500-736.

Harman, D. (2000) What we have learned and have not learned from TREC. *The BCS/IRSG 22nd Annual Colloquium on Information Retrieval Research*, Cambridge, 5-7 April.

Harter, Stephen, P. and Hert, Carol, A. (1997).Evaluation of information retrieval systems: approaches, issues, and methods. In: Martha E. Williams (Ed). *Annual Review of information Science and Technology* (ARIST), volume 32, 3-94.

Hearst, Marti. (1999) User Interfaces and visualisation. In: Baeza-Yates, Ricardo. and Ribeiro-Neto, Berthier. *Modern Information Retrieval*. Addison-Wesley.

Humphries, K.A. and Kelly, J.D. (1997). *Evaluation of search engines for finding medical information on the Web*. [Online] (URL: http://www.icaen.uiowa.edu/~humphrie/)

Information Retrieval Technology and Intelligent Concept Extraction™ Searching. (1996) [Online] (URL: www.excite.com/ice/tech.html) (visited 12 November 2000).

Jansen, Bernard J., Amanda Spink, Judy Bateman, and Tefko Saracevic. (1998) "Real Life Information Retrieval: a Study of User Queries on the Web. *SIGIR Forum* 32 No. 1 pp. 5-17.

Jansen, B.J. and Spink, A. (2000). Methodological approach in discovering user search patterns through web log analysis. *Bulletin American Society for Information Science*, 27(1). 15-17.

Kimmel, S. (1996) Robot-generated databases on the World Wide Web. *Database*, 19(1), 40-49.

Koll, M. (1993) Automatic relevance ranking: A searchers complement to indexing. *Proceedings of the 25th annual meeting of the American Society of Indexers*, p.55-60. Port Aransas, TX: American Society of Indexers.

Lancaster, F.W. (1979*). Infomation retrieval systems: characteristics, testing and evaluation.* 2nd edition. New York: John Wiley.

Lancaster, F.W. and Warner, A.J. (1993) Information retrieval today.
Arlington: Information Resources Press.

Large, A., Tedd, L.A. and Hartley, R.J. (1999). *Information seeking in the online age: Principles and practice*. London: Bowker Saur.

Larsen, R.L. (1997). Relaxing assumptions: stretching the vision. *D-Lib Magazine*, April 1997 [Online] (URL: www.dlib.org/april97/04larsen.html)

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.

Lesk, M.E and Salton, G. (1968). Relevance assessment and retrieval system evaluation. *Information Storage and Retrieval*, 4. 343-359

Meadow, Charles, T. (1986). Problems of Information Science research: An opinion paper. *Canadian Journal of Information Science*. 11, 18-23

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48, 810-832.

Nahl, D. (1998) Ethnography of novices' first use of Web search engines: affective control in cognitive processing. *Internet Reference Services Quarterly*, 32(2), 69.

Nasios, Y et al. (1998). *Evaluation of search engines*. Report undertaken by the National Technical University of Athens on behalf of the European Commission and Project PIPER, July 1998. Available at piper.ntua.gr/reports/searching/doc.0000.htm

Notess, Greg. (2000) Search engine showdown. www.notess.com/

Oppenheim, C., Morris, A., McKnight C., and Lowley S. (2000). The evaluation of www search engines *Journal of Documentation*, 56(2), 190-211.

Parasuraman, A., Zeithaml, V. and Berry, L.L. (1985) A conceptual model of service quality and its implications for future research. *Journal of Marketing*, 49 (Fall), p.41-50.

Parasuraman, A., Zeithaml, V. and Berry, L.L (1988) SERVQUAL: A multiple scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64 (1), p.12-40.

Robertson, Stephen, E. and Hancock-Beaulieu, Micheline, M. (1992). On the evaluation of IR systems. *Information Processing and Management*, 28(4), 457-466.

Salton, G and McGill M.J. (1983). *Introduction to modern information retrieval*. McGraw-Hill: New York.

Salton, Gerard. (1989*) Automatic text processing: the transformation, analysis and retrieval of information by computer*. Addison-Wesley, Reading, MA.

Salton, Gerard. (1992). The state of retrieval system evaluation. *Information Processing and Management*, 28(4), 441-449.

Sandore, Beth. (1990). Online searching: what measure satisfaction? *Library and Information Science Research*, 12, 33-54.

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.

Search Engine Watch. [Online] (URL: http://www.searchenginewatch.com/)

Sherman, Chris. (2000) *TheFireworksFly* .[Online] (URL*:* websearch.about.com/library/weekly/ aa041800b.htm)

Spink, A., Wilson, T., Ellis, D., and Ford, N. (1998). Modeling users successive searches in digital environments. *D-Lib Magazine*, April 98. [Online] (URL: www.dlib.org/dlib/april98/04spink.html)

Stobart, S. and Kerridge, S. (1996). *An investigation into World Wide Web search engine use from within the UK JISC-funded project undertaken by UKERNA*. [Online} (URL: http://osiris.sunderland.ac.uk/sst/se/results.html)

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.

Su, L. and Chen, H. (1999) *User evaluation of web search engines as prototype digital library retrieval tools*. CoLIS3, third international conference on conceptions of library and information science, Inter-University Centre Dubrovnik (IUC), Dubrovnik, Croatia, 23-26 May, 1999, 73-86.

Sullivan, D. (1998) Counting clicks and looking at links. *Search Engine Report* [Online]. (URL: www.searchenginewatch.com/sereport/9808-clicks.html)

Sullivan, Danny. (1999) Results Get More Targeted, *Search Engine Report* [Online]. (URL: www.searchenginewatch.com/sereport/99/04-excite.html)

Sullivan, D. (1999a) *How HotBot works* (Dec 1st, 1998). [Online]. (URL: http://searchenginewatch.com/subscribers/hotbot.html)

Sullivan, D. (1999b*) How Infoseek works* (Dec 1st, 1998). [Online]. (URL: http://searchenginewatch.com/subscribers/infoseek.html)

Sullivan, D. (1999d) AltaVista debuts search features. *Search engine report* (Nov. 4th , 1998). [Online]. (URL: http:// searchenginewatch.internet.com/sereport/9811-altavista.html

Sullivan, Danny. (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. [Online]. (URL: www.infonortics.com/searchengines/boston2000pro.html)

Sullivan, D. (2000a). Survey reveals search habits. *Search Engine Report*. [Online]. (URL: www.searchenginewatch.com/sereport/00.06-realnames.html)

Tessier, J., Crouch, W.W. and Atherton, P. (1977). New measures of user satisfaction with computer-based literature searches. *Special Libraries*, 68(11), 383-389.

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). Available at http://neal.ctstateu.edu:2001/htdocs/websearch.html

Travis, I. (1998) From 'storage and retrieval systems' to 'search engines': text retrieval in evolution. *ASIS Bulletin*, April/May, 2p.
Venditto, G. (1996). Search engine showdown: IW labs test seven Internet search tools. *Internet World*, May, 79-86.

Voorhees, E. and Garofolo, J. (2000) The TREC spoken document retrieval track. Bulletin of the American Society for Information Scientists, 26(5), 3p.

Wang, H., Xie, M. and Goh, T.N. (1999) Service quality of Internet search engines. *Journal of Information Science*, 25(6), 499-507.

Wiggins, R. and Matthews, J. (1998). Plateaus, Peaks and Promises: the Infonortics '98 search engine conference. *Searcher*, 6(6). Available at: http://www.infotoday.com/searcher/jun98/story4.htm

Wilder, R. *Evaluating search engines*. [Online] (URL:www.foley.gonzaga.edu/search.html)

Wiley, Deborah. (1998). Beyond Information Retrieval. *Database*, 21(4). Available at: http://www.onlineinc.com/database/DB1998/wiley8.html

Winship, I.R. (1995). World Wide Web searching tools – an evaluation. *VINE*, 99,49-54. Available at http://bubl.bath.ac.uk/BUBL/Iwinship.html

Zorn eta l (1999). Data Mining meets the web. *Online*, September/October, 17-28 [Online] (URL: www.onlineinc.com/onlinemag/)