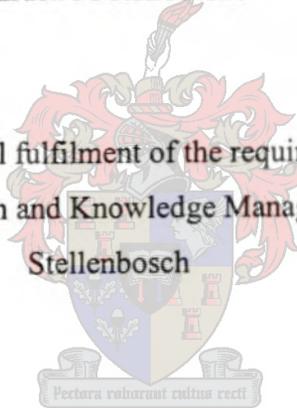


EVALUATION AND COMPARISON OF SEARCH ENGINES

Lindiwe Mtshontshi

Assignment presented in partial fulfilment of the requirements for the degree of
Master of Philosophy (Information and Knowledge Management) at the University of
Stellenbosch



Supervisor: Dr. M.S. van der Walt

December 2004

DECLARATION

I, the undersigned, hereby declare that the work contained in this assignment is my own work and that I have not previously in its entirety or in part submitted it at any university for a degree

Signature.

Date

Abstract

A growing body of studies is developing approaches to evaluate human interaction with Web search engines. Measuring the information retrieval effectiveness of World Wide Web search engines is costly because of the human relevance judgements involved. However, both for business enterprises and people it is important to know the most effective Web search engine, since such search engines help their users find a higher number of relevant Web pages with less effort. Furthermore, this information can be used for several practical purposes. This study does not attempt to describe all the currently available search engines, but provides a comparison of some, which are deemed to be among the most useful. It concentrates on search engines and their characteristics only. The goal is to help a new user get the most useful “hits” when using the various tools.

OPSOMMING

Al hoe meer studies word gedoen om benaderings te ontwikkel vir die evaluasie van menslike interaksie met Web-soekenjins. Om te meet hoe effektief 'n soekenjin inligting op die Wêreldwye Web kan opspoor, is duur vanweë die mens se relevansiebeoordeling wat daarby betrokke is. Dit is egter belangrik dat die bestuurders van sake-ondernemings en ander mense sal weet watter die mees doeltreffende soekenjins is, aangesien sulke soekenjins hulle gebruikers help om 'n hoër aantal relevante Webblaaie met minder inspanning te vind. Hierdie inligting kan ook gebruik word om 'n paar praktiese doelwitte te verwesenlik. Daar word nie gepoog om al die soekenjins wat tans beskikbaar is, te beskryf nie, maar sommige van die soekenjins wat as die nuttigste beskou word, word vergelyk. Daar word alleenlik op soekenjins en hulle kenmerke gekonsentreer. Die doel is om die nuwe gebruiker te help om die nuttigste inligting te verkry deur gebruik te maak van verskeie hulpmiddels.

Acknowledgement

Although written and attributed to one author, a dissertation is never merely the product of an individual effort, but rather the result of a variety of influences and personalities who contributed in a direct or indirect way to the final stage of the present study (for which the author, of course, takes the ultimate responsibility). However, I must confess that I do not have the space nor the time to catalogue all those whose contribution to this work is nonetheless immense, therefore I shall only mention a few who will represent all those I cannot mention.

First and far most I want to thank God the Almighty, who gave me the courage to start, the patience to continue, and the ability to complete.

I am especially indebted to Dr. Martin van der Walt, my supervisor for his kindness and constructive guidance.

I also wish to express my gratitude to Dr. Ben Fouche for all his effort in arranging for me to undertake this study and his willingness to help whenever I needed his professional guidance.

I would like to acknowledge my friend and colleague, Samuel, who despite his already busy schedule, patiently taught me computer skills.

I want to thank my parent, Mr Ezra MTSHONTSHI, for creating a desire in me to study further.

My thanks to Brownson Irondi for believing in me. To my biological mother, the woman so near and dear to my heart - I say maXhamela enkosi, and to my entire family at large, may God richly bless you all as we journey on.

Financial support from various sources is hereby acknowledged. Opinions expressed and conclusions reached in this study are those of the author and do not necessarily reflect their opinion.

Table of contents

Acknowledgement	iii
Abstract	iv
Table of contents	vi
List of tables	viii
List of figures	ix
1. INTRODUCTION	1
1.1 BACKGROUND	2
1.2 PROBLEM STATEMENT	4
1.3 OBJECTIVES	5
1.4 DELIMITATIONS	5
1.5 RESEARCH DESIGN AND METHODOLOGY	5
1.6 EVALUATION METHODOLOGY	8
1.7 OVERVIEW OF CHAPTERS	12
CHAPTER 2. RELATED STUDIES	13
CHAPTER 3. INTERNET SEARCH ENGINES:	16
3.1 WHAT IS A SEARCH ENGINE?	16
3.2 CLASSIFICATION OF SEARCH ENGINES	17
3.2.1. Directories:-	17
3.2.2. Search engines:-	17
3.3 SEARCH ENGINES' PROBLEMS	18
3.3.1. Time consuming and tedious:-	18
3.3.1. Relevance ranking:-	18
3.3.3. Query construction:-	18
3.3.4. Quick 'n Dirty vs. Continuous interest:-	19
3.4 THE COMMERCIALISATION OF SEARCH ENGINES	19
3.4.1 What should a searcher do?	21
CHAPTER 4. FACTORS AFFECTING SEARCH ENGINE PERFORMANCE	22
4.1 SEARCH SERVICE COVERAGE	22
4.2. INDEXING STRATEGIES	23
4.3. SEARCH FEATURES (USER CONTROL OF SEARCH)	23
4.4. USERS AND USAGE OF SEARCH ENGINES	24
4.5 SEARCH FEATURES (SYSTEM CONTROL OF QUERY)	26
4.5.1 Query expansion	26
4.5.2. Query modification	27
4.5.3 Query visualisation	28
4.5.4 Popular queries	29
4.6 RESULTS DISPLAY	30
4.6.1 Ranking	31
4.7 CHAPTER SUMMARY	33

CHAPTER 5.INDIVIDUAL SEARCH ENGINES	34
5.1. ALTAVISTA	34
5.1.1 Simple searches	35
5.1.2. Advanced searches	37
5.1.2.1. Boolean and proximity searching:	37
5.2 HOTBOT	40
5.3 EXCITE	42
5.4. LYCOS	44
5.5 WEBCRAWLER	46
5.6 YAHOO	48
5.7 INFOSEEK	50
5.8 OPENTEXT	52
CHAPTER 6	54
6.1 DISCUSSION AND CONCLUSION	54
6.2 RECOMMENDATIONS - Choosing the search engine	56
6.3. RECOMMENDED SEARCH ENGINE	58
REFERENCES	59

List of tables

Pages

Table 1. Task model, based on a simplified model of the information access process-	5
Table 2. Comparison of evaluation criteria and system/context parameters -----	10
Table 3. Search engine coverage -----	17
Table 4. Search engine indexed terms -----	18
Table 5. Search engine and search features -----	19
Table 6. Search engine search features (system control) -----	25
Table 7. Search engines results displayed -----	26
Table 8. Search engine ranking boost -----	28

List of figures	Pages
Figure 1 AltaVista -----	32
Figure 2 HotBot -----	38
Figure 3 Excite -----	40
Figure 4 Lycos -----	42
Figure 5 WebCrawler -----	44
Figure 6 Yahoo -----	46
Figure 7 Open Text -----	49

1. INTRODUCTION

We as professionals, do not use every search engine or Web directory daily, nevertheless, we have to know how each works and what data each does or does not contain. I fully understand that this is easier said than done, but today information access is a topic that everyone is aware of and talking about. Pick up any newspaper. Turn on the television. Everyday more and more articles and reports discuss searching the Web. Many of these articles and reports are written for and by non-information professionals. We have to stay ahead of our clients and patrons if we hope to help them. Excite or AllTheWeb may not be your search engines of choice, but quite possibly, they are for someone you know. Our colleagues, co-workers, and friends come to us as the "search experts" and we must do our best to help. Our knowledge and understanding in this area are great ways to make our profession look good and to make our already valuable jobs even more valuable.

The Internet is a worldwide network of computers connected to each other through telephone lines, satellites, fibre optic cable and other means. These connections allow users to exchange and share information via computer through an Internet connection. The World Wide Web is that part of the Internet containing text, graphics, sounds, movies, and more. There is a vast amount of information available on the Internet, the good, the bad, and the ugly.

Today searching on the World Wide Web (WWW) seems to have become part of our routine life. The Web has even become an essential tool for collecting information. Undoubtedly, the Web provides the convenience, but this sea of information makes any query on this huge information reservoir extremely difficult, so that people think it is a chaotic repository (Lynch 1997). In order to solve this problem, some companies have developed search aids (search tools) on the internet such as Yahoo, InfoSeek, AltaVista, Lycos and many more. Obviously, you cannot use all of them at the same time. Faced with so many search tools, people can get confused quite easily. Which one is the best? Which one should I use? The present study tries to answer these questions by comparing eight search tools: AltaVista, Lycos, Hotbot, Yahoo, InfoSeek ultra, Excite, WebCrawler and Open Text

1.1 BACKGROUND

We are a society obsessed with convenience. We go to extremes to invent devices that promise a simpler and more convenient lifestyle. This paradox is exemplified in our fascination with the internet, as well as with our attempts to index it for access purposes. The internet is a rapidly evolving organism that is almost completely lacking in fundamental organisation. The question whether each individual achieves a net gain from all the effort expended in this process lies somewhere beyond the scope of my project, but I think we all can agree on the need to organise this much unstructured information resource somehow.

Internet search tools have been created to answer this very pressing need. They are evolving rapidly, some would say more rapidly than the internet itself. By the end of 2005, it is estimated that the internet will consist of no less than 60000 million pages, containing 70 or 80 billion words. To make matters worse, this great mass of data exists completely without any kind of bibliographic control, standard numbering, or classification system. Clearly, automated tools of some sort are necessary to sift through this mass of material (Venditto, 1996).

Internet search engines have proliferated with the growth of the Internet itself. There is a growing number of major general purpose search engines from both established commercial firms in the industry and from new 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 one of development and competition. Advances in search engine technology are reviewed annually at the Infonortics Search Engine Meeting (Wiggins and Matthews, 1998; Wiley, 1998; Sullivan, 2000).

Search engines are often categorised as robot-driven devices which respond to a user query or as directory-based systems that guide users through classification 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 Real Names (Sullivan 2000) 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.

In general Internet search services are built using 'spiders' or software programmes to create and maintain a proprietary index of Web documents, search engines, 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, and sustainable and use a variety of techniques derived from 30 years of research in IR to achieve their performance levels. However, considerable variations exist 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 others. 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 will be able to fulfil the users' expectations for finding precise information.

However as Evans (in Wiggins and Matthews, 1998) have noted, search engine developers may be approaching a fundamental limit in terms of the capabilities of their systems. Evans describes the 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". Feldman (1998) warns that the problem facing the developers is more fundamental; stating that the Web searching market is 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 uncertain.

The uncertainty that surrounds users' expectation and usage of search engines give rise to the question of how we can evaluate the impact of their development 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.

My own personal interest in the internet has increased in direct proportion to the growth, power, and flexibility of the excellent search tools that have appeared over the past year or two. In my opinion, they have elevated the Web from a simple "browser paradise" to a more respectable, searchable, and interesting World Wide reference source.

1.2 PROBLEM STATEMENT

This paper attempts to compare and evaluate Web index services in an objective and fair manner. One begins such a study by choosing which services to compare. One must develop an unbiased suite of queries to test those services. Then one must design fair methods for searching the services and evaluating the results of the searches. At many points in the design, it is possible to subtly favour one service over another. Conscious or unconscious bias must be guarded against.

Here we chose our services--Alta Vista, Excite, Hotbot, Open Text, Infoseek, Lycos, Yahoo and WebCrawler. Other service that could have been chosen in the category of major search service is Google. Time did not allow for its inclusion.

There is explosive growth of information on the World Wide Web, which poses a challenge to traditional information retrieval (IR) research. Other than the sheer amount of information, some structural factors make searching for relevant and quality information on the Web a formidable task. The free wheeling nature of publishing on the Web is a blessing for the flow of ideas, but it has also complicated the process of retrieving relevant information. In contrast to traditional IR, there are no consistent indexing and classification principles for organising material on the Web. Nor are there any filtering practices at hand to ensure the quality of and credibility of the documents. Furthermore, certain features of the Web search situations also distinguish the Web from the traditional IR setting. It has been shown that ordinary Web searchers tend to give little input (Leighton 1997) and are very

sensitive to the time and effort put into the search (Wishard, 1998). The issues of credibility and user efforts peculiar to the Web search environment are not addressed properly by traditional precision and recall measures. Several measures that focus on user efforts have been proposed, yet there has been little investigation on the validity of these measures.

1.3 OBJECTIVES

The overall aim of this project is not to describe all currently available search engines but to provide a comparison of some, which are deemed to be among the most useful. It concentrates on search engines and their characteristics only. The goal is to help a new user get the most useful “hits” when using the various tools.

1.4 DELIMITATIONS

Delimitations set the boundaries for a study. This paper evaluates eight search engines that are not commonly used, as compared to Google.

1.5 RESEARCH DESIGN AND METHODOLOGY

Web search evaluation poses a considerable number of challenges to traditional IR evaluation methods. First the collection is constantly changing i.e. any evaluation is not reproducible in future. Since the collection is so large, it is not possible to judge enough queries manually and to a sufficient result depth to be able to measure recall in any reasonable way.

This paper reports results from an exploratory study evaluating these search engines: 1.AltaVista, 2.HotBot, 3.Excite, 4.InfoSeek ultra, 5.Lycos, 6. WebCrawler, 7.Yahoo 8.OpenText.

The evaluation approach explored in this study is based on the user-tred approach discussed by Spink et al (1998), who proposed that search engine evaluation should

focus on measuring the impact of users' interactions on their information problem and their moves through the different stages of their information seeking process. Below is the task model, based on a simplified model of the information access process.

The specific task domain is users' wish to retrieve relevant items to satisfy their information needs. Although individuals' information seeking goals can differ quite widely, standard models of the information seeking process contain the core steps of query perfection, receipt of results in an interaction cycle. The process on which the researcher draws identifies interacting steps, which are not necessarily sequential and may be repeated. This gives dimensions on which users might evaluate their satisfaction with the system. These are:

1. User will formulate or submit a query
2. User will receive results
3. User will evaluate results – end or modify (note possible feedback loop here)
and
4. User will evaluate success of the search as a whole.

For each dimension, we can relate the criteria of Effectiveness, Efficiency, Utility and Interaction, by which the user might evaluate his/her satisfaction with the system on these task dimensions.

Dimension 1 User will formulate/submit a query evaluated on the criterion of interaction (query)

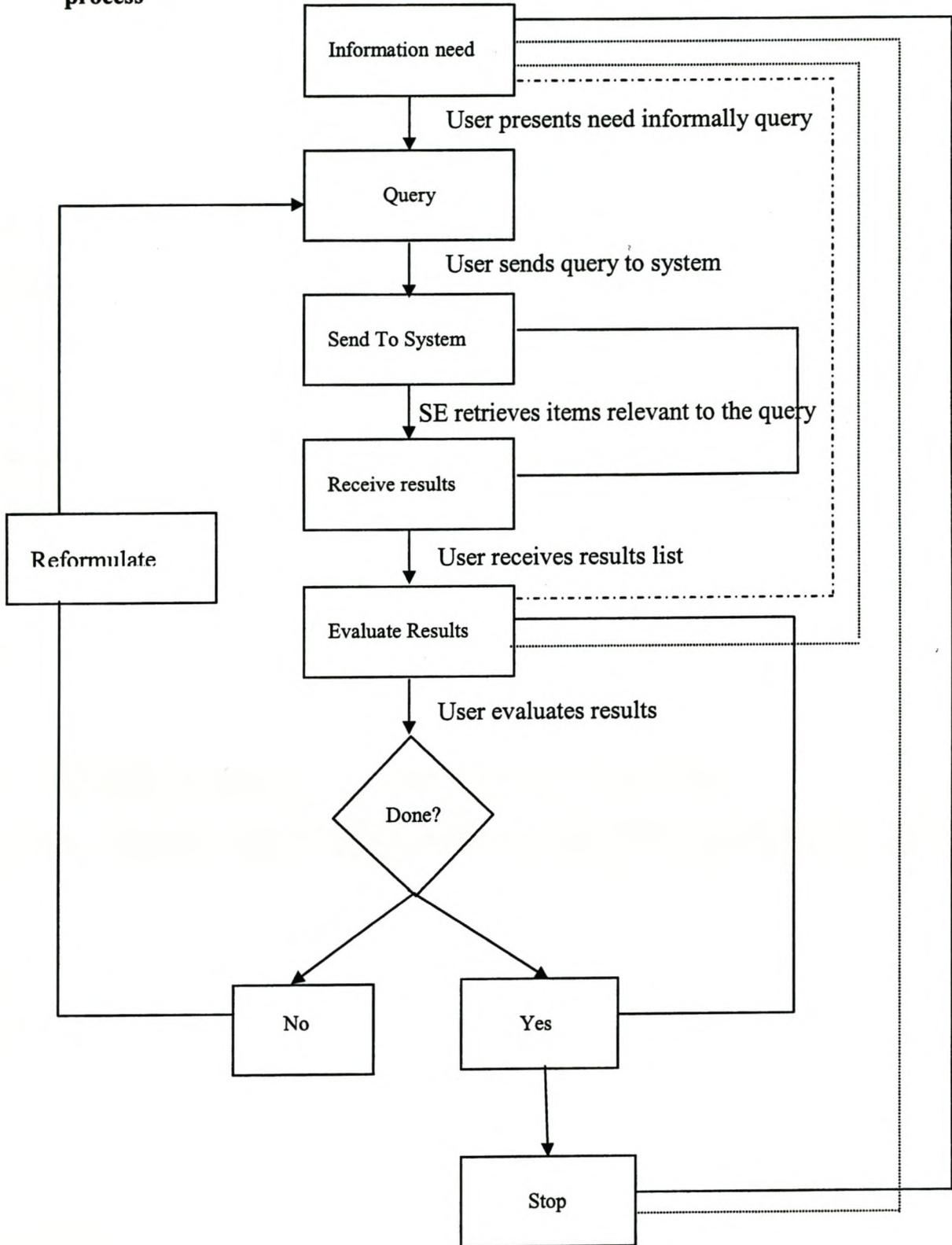
Dimensions 2 User will receive results evaluated on the criterion of interaction (output)

Dimension 3 User will evaluate results evaluated on criteria of effectiveness & relevance/ranking

Dimension 4 User will evaluate success of the search evaluated on criteria of efficiency & utility

It is important to note that this model has been contrasted with others such as Bates' (1998) 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 is acquired along the way.

Table 1. Task model, based on a simplified model of the information access process



1.6 EVALUATION METHODOLOGY

Evaluation is a process by which the effectiveness of a 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, Tedd & Hartley, 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 2:-

- 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 by 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 evaluation 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 the study focuses 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, the users' interface may affect user satisfaction and thus have an important role to play in users' selection of systems. Further, the table sets the 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 technique and retrieval algorithms could impact on retrieval performance.

These parameters become increasingly complex as the measure becomes more user orientated as not only do they define what is evaluated but equally the parameter's 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 users' 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 for 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 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's use of search features to formulate a query statement, facilitating effort in narrowing a search. Indeed, the interface (and non-retrieval devices) may affect the whole mode interaction for the user and hence influences 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 user's task may also influence their search behaviour.

Table 2 Comparison of evaluation criteria and system/context parameters

Evaluation criteria		System parameters	User context
1. Coverage (proportion of literature on the topic)	<ul style="list-style-type: none"> • Composition of Web indexes • Coverage using the Clarke and Willet method 	Composition of the index will affect the performance of the search engine	
2. Recall (retrieve relevant items) Precision (hold back non relevant items) <ul style="list-style-type: none"> • RELEVANCE (precision, relative recall, user vs. system ranking) 	<ul style="list-style-type: none"> • Retrieval based on precision • Performance based on precision and relative recall 	The indexing language, exhaustively and specificity, and retrieval mechanism will affect performance	Query formulation and search strategy
3. Responsive time (from request to results) <ul style="list-style-type: none"> • EFFICIENCY (search session time, relevance assessment time, cost) • USER SATISFACTION with response time 	<ul style="list-style-type: none"> • Response time • Response time 	As above, and organisation of stored documents, size of collection, file format will affect response time	As above, type of query

4. 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
5. Format (presentation of 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 of results 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
6. User effort (expended to achieve a satisfactory response) USER SATISFACTION with search interface and online documents	User effort based on analysis of documents 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

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 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 successful if users' ratings on various scales of satisfaction are at a maximum. Research into users and their 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, Zeithaml & Berry, 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 expressed 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.

1.7 OVERVIEW OF CHAPTERS

Chapter 1 provides a general introduction, background, problem statement, objectives, delimitations, research design and methodology to search engines and their evaluation.

Chapter 2 presents the related studies to the evaluation and comparison of search engines.

Chapter 3 explains the categorisation of search engines and their features

Chapter 4 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 users 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.

Chapter 5 contains the exact results on evaluation and comparison of seven search engines and are displayed and explained.

Chapter 6 concludes the whole matter and states the recommendations.

CHAPTER 2. RELATED STUDIES

Since search engines came out only in 1994 (Chu and Rosenthal, 1996), there was not much prior comparison associated with the performance of these search tools. The comparisons available were descriptive reviews that were highly dependent on individual experience, thus the results are varied and some of the reported findings do not appear to agree with one another. These evaluations were performed according to some of the main factors that determine the success of a search engine, such as the size, content of the database, the speed of searching, update frequency, the availability of search features, the interface design and the ease of use (Willis 1996, MacDonald 1996, Richard 1997. Lynch 1997)

Obviously, in order to be effective on the Web, it is important to utilize the search engine most suited to people's subject domain. The following reviews neither include a ranking, which could help to make a decision for one specific search engine, nor did some of the rankings provided have a scientific basis. Leighton (1996) considered this problem and used eight reference questions from a university library as search queries. By employing the evaluation criterion precision, he compared Info seek, Lycos, WebCrawler and World Wide Worm. However, he counted only the number of relevant links. Westera (1996) only used five queries to conduct a precision study, and all these queries dealt with wine. These test suites were too small for statistical usefulness. Chu and Rosenthal (1996) studied first ten queries to conduct precisions (ratio of retrieved relevant documents to the total number of relevant documents in the database), took enough queries for statistical comparisons, recorded crucial information about how the searching was conducted and performed some statistical tests. However, they did not subject their mean precision figures to any test for significance.

Schilichting (1996) tested first ten queries conducted on precision and listed query topics that he and his helpers searched. However, they used structured search expressions (using operators) and did not list the exact expression entered for each service. They reported the mean precision, but again did not test for significance in differences. They even did not have criteria for relevance.

Leighton (1997) conducted a new study for his Master's thesis to correct the problems presented in his early study in 1995. He tested the precision of the first twenty results returned for fifteen queries and used Friedman's randomised block design to perform multiple comparisons for significance. To avoid the question of bias, he used a blinding algorithm for evaluator to know from which search service the citation came by developing a PERL program. Clearly, it is impossible for a user to conduct the general evaluation by him-/herself as it will take several months.

Another interesting study conducted by Schlichting and Nilsen (1996) evaluated AltaVista, Excite, Info seek and Lycos. They applied signal detection theory to analyse sensitivity and how conservative or risky the search engines (beta) were for finding useful information, but they did not conduct significance tests.

Several studies have explored the applicability of traditional IR evaluation i.e. precision and recall, on search engine performance (Wishard 1998, Clarke and Willet 1997). Leighton studied the precision of ten queries on three search engines. Instead of a binary measure of relevance (relevant/non-relevant), they adopted a three point scale to distinguish among relevant, partially relevant and non-relevant documents. Clarke and Willet also used a three-point scale in assigning relevance scores, with a slight modification: pages that were considered non-relevant in themselves, but that led to relevant pages, were judged partially relevant. In the absence of a predefined set of relevant documents, Clarke and Willet found it very difficult to assess recall on the Web. Clarke and Willet constructed a relative recall measure by using the merged outputs of all four search engines tested as a pool of relevant documents.

Some characteristics of Web searching, however, require performance criteria other than the precision and recall measures developed in traditional IR. The enormous amount of information and the wide variety of sources of the Web seem to make quality of ranking a much more important dimension in assessing search engine performance, since users in general spend less time and effort to sort through the retrieved pages. This is supported by studies of users' searching behaviour on the Web.

Silverstein et al (1998) found that 85 percent of users look only at the first screen with results. They applied five criteria – relevance, efficiency, utility, user satisfaction, and connectivity to evaluate the performance of four search engines. Furthermore, instead of submitting simple text queries as in most search engine evaluations, they used real user search strategies and judgement in the searching and evaluating process.

In contrast to traditional IR searchers the majority of Web users are laypersons who are more sensitive to time and effort spent on finding information. The ability to optimize search order thus becomes an even more salient dimension of search engine performance.

The notion of Expected Search Length (ESL) first proposed by Cooper (1968) seems to be an ideal notion to test whether a search engine is able to deliver the most relevant documents at the top of retrieved sets. According to Cooper, the primary function of a retrieval system is to save users as much labour as possible in the search for relevant documents, by perusing and discarding irrelevant results.

In summary, there is no generally effective methodology to compare and thus evaluate search engines so far.

CHAPTER 3. INTERNET SEARCH ENGINES:

The categorisation of search engines 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. I 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 characteristics of engines by features which provide clues for the development of an evaluation methodology.

3.1 WHAT IS A SEARCH ENGINE?

Before we begin, we need to get a definition straight — a definition that I think many of us has thought about. What does "Web search" mean to the information professional? In the early days of the Web, it meant exactly what it sounds like — material found on the open Web.

However, as we move forward, the term "Web search" has taken on new meanings. Does a Web search involve tools like Google or AltaVista to reach "open access" material? Does it mean using the Web as a vehicle to log onto proprietary databases such as Factiva or Dialog? Not too long ago, logging onto proprietary services required individual connections to each one. Today, any Web browser with an Internet connection can reach those services. Perhaps it means both. This lack of common understanding can confuse some.

A search engine as defined by Gordon and Pathack (1999) as a logical formula that uses a computer program to compare your search to items in the index or database to produce your result. Your choice of a search engine should be based on the search that you wish to do, and the functionality of the search engine itself. The capabilities of a particular search engine will determine how you construct your search.

3.2 CLASSIFICATION OF SEARCH ENGINES

To understand how a search tool works, it is necessary to know the classification of current tools. Demoss (1996) divides the search aids into the four categories of search engines, directories, software search and all-in-one-search. According to how a search aid works, other researchers classified these services into two basic types: Directory and Search Engine (Lui 1996, Richard 1997) which are briefly introduced below.

3.2.1. Directories:-

A directory can be regarded as a manual catalogue of sites on the internet. This means that people actually create categories and assign sites to a place within a structured index. One typical directory of this kind is Yahoo, the staff of which screens all relevant information and assigns this information to its relevant address. Yahoo also orders sites so that the most relevant or comprehensive in each category appears first on the list. This search feature can help people quickly find targeted information or topics that are more general.

3.2.2. Search engines:-

Actually, the main difference between directories and search engines is that a directory is built by people, but the search engine's database is created by software known as spiders or robots. This is why some people call search engines robot-driven or robot wanderer, or spider, harvest and many more (Liu, 1996; Westera 1996). These spiders attempt to crawl through the Web, collect and index resources in an automatic fashion, and put this information into a database by using their own specific algorithm. Thus they do not need extensive human intervention. Searching, instead of browsing, is the main feature of this type of tool. The other component is the query module. Users search the index through a predefined query module, an interface specific to each engine.

The advantage of search engines is that they are nearly comprehensive, often including thousands of sites in the results listed. Certainly, it is also useful when you are searching for a specific topic that may not be found in a directory. The

disadvantage is that you often have to weed through piles of non-relevant sites to find what you are looking for. This study is designed to determine which of several interactive search tools is best in terms of usability and the quality of retrieved Web sites for various queries.

3.3 SEARCH ENGINES' PROBLEMS

3.3.1. Time consuming and tedious:-

Measuring the search engine's effectiveness is expensive due to the human labour involved in judging relevancy. (For example, one subject of the experiments spent about 12 hours to judge the query results) Evaluation of search engines may need to be done often due to changing needs of users or the dynamic nature of search engines (e.g. their changing Web coverage and ranking technology) and therefore the evaluation needs to be efficient (Hawking, Craswel, Bailey, & Griffiths, 2001). Query refinement is often necessary, several sources must be used, and area monitoring means repetitive searching.

3.3.1 Relevance ranking:-

For assessing the performance of search engines, there are various measures such as database coverage, query response time, user effort, and retrieval effectiveness. The most common effectiveness measures are **precision** (ratio of retrieved relevant documents to the total number of relevant documents in the database) and **recall** (ratio of retrieved relevant documents to the total number of relevant documents in the database). These are difficult to understand and difficult to visualise.

3.3.3. Query construction:-

People use search engines for finding information on the Web. A Web search engine is an information retrieval system (Salton & McGill 1993) which is used to locate the Web pages relevant to user queries. A Web search engine contains indexing, storage, query processing, spider (or crawler, robot), and user interface subsystems. The

indexing subsystem aims to capture the information content of Web pages by using their words. During indexing, frequent words (that, the, this, etc.) known as stop words, may be eliminated since such words usually have no information value. Various statistics about words (e.g. number of occurrences in the individual pages or in all of the indexed Web pages) are usually stored in an inverted file structure. This organisation is used during query processing to rank the pages according to their relevance scores for a given query. Hyperlink structure information about Web pages is also used for page ranking (Brin & Page 1998; Koboyashi & Takeda, 2000) the spider subsystems bring the pages to be indexed to the system

However, Web users think of a search engine as nothing but its user interface that accepts queries and presents the search results.

3.3.4. Quick 'n Dirty vs. Continuous interest:-

There are millions of Web users and about 85% of them use search engines to locate information on the Web (Koboyashi & Takeda, 2000). It has been determined that the use of search engines is the second most popular Internet activity next to e-mail (Jansen & Pooch, 2001). Due to high demand, there are hundreds of general purpose and thousands of specialised search engines (Lawrence and Giles, 1999). Search engines are good in the case of "precise questions" and lousy for vaguely defined "areas of interest".

3.4 THE COMMERCIALISATION OF SEARCH ENGINES

This issue has received a great deal of well-deserved attention lately. It seems to me that the wants and needs of the searcher/researcher and the many people from various groups (the engines themselves, the search optimization community, and the advertising community) have different ideas about what the bottom line is when it comes to Web searches. Don't misunderstand me — the engines are profit-making-businesses, or try to be, so making money is goal number one. I understand this fact. However, those of us who use the "open Web" as a research tool want timely and

authoritative answers without advertising or undue influence getting in the way of the best possible answer available.

Can the wants and needs of the two groups co-exist? Absolutely, but it will take knowledge and continuing education for both information professionals and end users to continue to use general-purpose Web search tools as effective resources. The bottom line here is knowledge of the issues for all parties. Using the Web effectively without general-purpose search engines would be difficult, time consuming, and in many cases impossible. This is particularly true for the professional researcher.

Pay-per-placement, pay-per-click allows a person or company to buy a keyword or keywords and have their results at the top of the results list when a word or words are searched. GoTo.Com is just one of many examples of this type of search engine. The extra challenge with GoTo and others is that in addition to searching at GoTo.Com they also sell their database to other engines for them to brand as their own. For example, GoTo.Com "powers" NBCi and Go.Com (formerly Infoseek). So, if a user tells you that NBCi is his or her engine of choice, in actuality they are searching GoTo.Com material. Various "flavours" of this type of branding exist in the Web search world. To get an idea of how many of these engines are online check <http://www.payperclicksearchengines.com>.

Paid-inclusion programs available from many of the leading engines have programs in place that will allow a person or company to pay a fee and make sure that their site is crawled and included in that particular database. Additionally, this fee will also make sure that the site is recrawled on a regular basis, sometimes every week or so. This can mean that searchers may assume a currency of results based on retrieval from the paid-inclusion sites that does not occur with non-paying sites.

Search optimization consultant's reverse-engineer search engines and relevancy-ranking algorithms and then use this knowledge to place a client's Web pages higher in a search result list.

Danny Sullivan, the editor of Search Engine Watch [<http://Searchenginewatch.com>], covers this and most other parts of the search world on a regular basis and at great

depth. Also, to learn more about search engine optimization take a look at Rank Write Roundtable [<http://www.rankwrite.com>]. By the way, keeping current with the search engine optimization discussion can often provide searchers with deep background about how the engines work. Again, this makes for a better searcher.

3.4.1 What should a searcher do?

Understand the differences between search engines, become familiar with the terminology, and share this knowledge with others.

In the case of more "traditional" engines, be aware of how commercial material is labelled and where it is placed. For example, AltaVista offers "partner listings" at the top and bottom of a results list. Excite uses the term "sponsored link." Hotbot places "products and services" at the top of the results list.

CHAPTER 4. FACTORS AFFECTING SEARCH ENGINE PERFORMANCE

4.1 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 with agents or robots. Table 3 below 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. If an engine does not support the use of frames and image-maps, this 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 use “learn frequency” to re-examine sites which change frequently. Instant index refers to the time delay for trawled pages to appear in the index. While the process of selecting and/or reviewing quality content is generally reserved for subject-specialised search services, also attempt to reduce the size of the database by establishing subsets of reviewed resources or the 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 to it from other pages.

Table 3. Search engine coverage

Coverage	AltaVista	Excite	HotBot	Infoseek	Lycos	WebCrawler	Yahoo	Open Text
Estimated size	30m	50						
Deep crawl	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Frames supported	Yes	No	No	Yes	Yes	Yes	Yes	Yes
Image maps	Yes	No	No	Yes	Yes	Yes	Yes	Yes
Learns Frequency	Yes	No	No	No	No	Yes	Yes	
Instant Index	Yes	No	No	No	No	No	Yes	
Coverage content	www	www & reviewed sites	www	www	www	www	www	www
Link popularity	No	No	Yes	Yes	Yes	Yes	Yes	Yes

4.2. INDEXING STRATEGIES

The list of 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. Stop words may or 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, has seemingly been discarded by search engines. Web site developers have reportedly misused metatags, for example repeating terms many times, in the attempt to have a page appear in the top ten retrieved. HotBot reportedly enhances its index with human intellectual representations of items. Some services offer a combination of catalogs (selected collections described and classified into 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 4. Search Engines’ indexed elements

Indexing	AltaVista	Excite	HotBot	Infoseek ultra	Open Text	WebCrawler	Lycos	Yahoo
Full text	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stopwords/not searched	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Meta descriptions	Yes	Yes						
Meta Keywords	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Comments			Yes					
Subject categories								

4.3. SEARCH FEATURES (USER CONTROL OF SEARCH)

The graphical user interface, GUI of a search engine provides systems with a mechanical device whereby the control for interaction is placed with either the system or the user. AltaVista for example provides the option of simple querying, advance query, and predefined category browsing. In the opening screen (typically the simple query mode) most search engine interfaces focus on supporting the user’s 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

guidelines as to how to enforce the processing of the query terms (match all / match any / treat as exact phrase / include or exclude 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 arrays of more advanced search capabilities are shown in the table. The use of these, for example Boolean, to specify query term relationships and truncate 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 5. Search engines' search features

Search	AltaVista	Excite	HotBot	Infoseek ultra	Lycos	Open Text	WebCrawler	Yahoo
Boolean search	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nested parenthesis	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Include/Exclude (+ -)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Default	OR	OR	OR	AND	AND	AND	AND	
Proximity/near/adjacency searching	Within 10 words	Concept search approximates this	no	Relevant ranking gives boost for nearness				
Phrase search	Yes	Yes	Yes	Yes	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	Yes	Yes	Yes	
Case sensitivity (wholly, partially) Field search e.g. based on title text, site, url, link, host, domain, anchor, image)	Yes	No	For a person search yes	Will boost search if capitals in results when used in query Yes	No	No	No	No
Limit restrictions (e.g. based on date, language, subject, document type, industry, domain, etc)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

4.4. USERS AND USAGE OF SEARCH ENGINES

Search engines offer an array of search features found in traditional online services. Yet whilst many of those features give a trained search intermediary optimal search

performance, search engines users are likely to range from expert to casual (Travis, 1998). Wiggins and Matthews (1998) in summarising the themes of the 1998 Infornotics conference highlighted the consensus which was the driving force behind many of the developments reported. Professional searchers may be adept at using the 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 description 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.

A survey of Web usage gives some indication of what the average Web searcher is doing and points to differences between Web searches and queries with traditional IR systems. Observation of average Web search (Spink et al, 1998 Elis et al 1999) point out that their ineffective use may be owing to the poor understanding most users have as to how a search engine automatically discriminates between single terms and phrases. Few are aware when a search service defaults to AND or NOT, and expect a search engine to discriminate automatically between the single term and phrases. Further devices such as relevance feedback, seemingly conducive to end-users' 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 of 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 the 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 that they are looking for. When adopting the – *I will 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 prototype 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 keywords, 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 centre 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 is diluted by 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.

Thus in the context of categorising the development of search features, we distinguish those which provide searchers with assistance and those which shift the control back to the system to provide the most likely relevant hits.

4.5 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 do not 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 4, For example where Infoseek Ultra will boost the ranking of retrieved items containing capitalised 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 directions to help the user browse more productively.

4.5.1 Query expansion

Help in improving a user query formulation may be provided by the use of concept searching. The assumption here is that users will take a quick and simple approach to putting a query to a search 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 ICE technology)

Automatic query expansion uses a system-generated thesaurus, better 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 a high recall rate typical of a thesaurus-based system and, since precision can be 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 with Latane Semantic Indexing which analyses, by correlation 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 the 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' or 'bottles' then, it is likely to refer to a room containing a counter across which refreshments are served rather than a rod, a place in court behind which a prisoner stands. Furthermore, if these clusters of symbols are present in a text, there is a good chance that it is about the said concepts even if the word 'bar' is not actually present. As far as the user is 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 the other, but which often result in little or no overlap in the results obtained when put to the same search engine.

4.5.2. Query modification

Providing more user control 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 categorization of documents by topics is likewise offered as a browsable suggestion of

topics likely to be relevant to a given search. All these may assist the user in narrowing a search and provide more precision in the search results.

Another technique 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 the user in control of the interaction. The basic principle is that users control subsequent queries by assessing the relevance of documents, which is then used to modify the 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.

4.5.3 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 search results. That is the creation of on-the-fly groupings of search results which 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 user in judging the relevancy of the retrieved items. It also has the useful side effect of highlighting to the user potential ambiguity of the original search terms (as has been noted, users often fail to provide the important contextual information of the query) and thus can be viewed as query assistance. Excite's ICE technology recognises clusters of documents and can use this as the basis for the grouping of the search results. The most elegant are the Infoseek dynamic custom folders (Zorn et al, 1999) which are based on the

categorisation of documents in which documents are mapped to a classification system and tagged accordingly. Custom folders based on the search results set provide the user with a hierarchical overview of the major topics retrieved, allowing the drilling down from the broad to the specific and aiding the browsing of different themes or concepts within search results.

4.5.4 Popular queries

Search assistance can thus be provided in the form of query expansion, and query modification or visualisation of the major topic 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 used terms and specification of relationship among query terms. Most traditional information retrieval techniques rarely deal with further complexity in the way in which humans are accustomed to conveying the meaning of understanding discourse. Much of what we convey is in what is not said and is 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 outcome, prior knowledge of what users are searching for, can be factored into the retrieval strategies of search engines.

"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 real link which will direct a user to official sites when a brand name search is conducted. Hotbot's related searches offer searches which are similar, either more general or more specific, to a given query. Excite's "target results" responds to certain geographical locations such as 'Cape Town' and

will offer first its list of pre-programmed results of custom information including a city map, tourism resources, current weather, etc. In a sense the search engine presumes that this is the type of information the user is likely to be searching for when entering a general query

Table 6. Search engine search features (system control)

Search features (system control)	AltaVista	Excite	HotBot	Infoseek	Yahoo	WebCrawler	Lycos	Open Text
Query expansion		Concept search		Concept processing?				
Query modification	Refine (suggest terms)	Search wizard				Search Wizard (suggest terms)		Handle plurals
Query visualisation		Cluster group search results					Common folders	
Popular queries	Related searches	Related searches	Related searches		Real Names Related searches			Related searches

4.6 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 retrieval items 10 to a page in a default format showing at least a minimum of results and some text. Format displays can usually be changed with commands such as: Sort by date, cluster by site or sort by URL (to identify pages from the same site and thus prevent any one site from dominating the results). The summary may vary in size and preparation, e.g. some are pre-prepared or automatically constructed or use 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 7. Search engines results display

Display	AltaVista	Excite	HotBot	Infoseek ultra	Lycos	Open Text	WebCrawler	Yahoo
Sort by options	No	Yes	No. but offers clustering	Yes	Yes	Yes	yes	Yes
Results at time	10	10	10	10	10	10	10	10
Title size	78	70	80	80				
Summary size	150	395	170-250	150-200	200			
Metatags description	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Highlight search terms	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

4.6.1 Ranking

In terms of judging the results list Courtois and Barry (1999) argue that users are most likely to scan their results list and select some items only. 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 and 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.

Courtios and Barry (1999) point out that popularity of search engines is due in part to the perceived ease of use caused by their use of ranks output. The results and their relevancy to a given query are usually ranked by statistical term frequency, location, and possible 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 1999). 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 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 (1999) 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 Intelligence Concepts (Excite 2002).

These methods for ranking output relevance have been experimented with for decades, but are limited to relevance based on topic alone. Barry and Schamber (1998) list 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 or ranking from a user point of view, to minimize the effort of searching for an item, search engines have adopted a variety of experimental approaches using off-the-page parameters to boost the ranking of an item.

Links popularity boosts the ranking of a site - if many other documents link to it, it is considered to be popular. Generally speaking, counting links will place those with most points 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 capitalise automatically on the human endorsements of Web pages made by site authors. Direct Hit is a company which works with search engines (e.g. HotBot) and monitors users' clicks on search results (what page 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. When seeking information with a search engine, the user may be offered the Direct Hit. For example in HotBot Direct Hit results are displayed under the heading “Web search: top 10”. This is usually available 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. Received 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 boost ranking if a search term appears in a metatag.

Table 8. Search engines ranking boost

Ranking boost	AltaVista	Excite	HotBot	Infoseek Ultra	Open Text	WebCrawler	Yahoo	Lycos
Link popularity	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Direct Hit	No	No	Yes	No	Yes	Yes	Yes	Yes
Reviewed status	No	No	No	no	Yes	Yes	Yes	Yes
Metatags	no	no	Yes	no	no	no	Yes	Yes

4.7 CHAPTER SUMMARY

The review has presented a very broad categorisation of search engine components to show the extent of variation of features in individual search engines which may impact on their performance. Any combination of these may lead to a more effective search, and thus improve 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. It could be argued, as in my introduction, that search engine developers are targeting a niche, a type of user and/or information query. Furthermore 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 the content of large collections or search results, the latter is used to recommend documents by analysis of citation paths or hyperlink paths. It would appear that the shift towards supporting users in their information seeking tasks, possibly to the extent of providing the information even if it was not requested, would 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. It is not known what will satisfy users and what impact these more novel features might have on search satisfaction.

CHAPTER 5.INDIVIDUAL SEARCH ENGINES

5.1. ALTAVISTA

URL: <http://www.Altavista.com>



The AltaVista search revolution started in April 1995. Digital had just released Alpha 8400, a 64-bit monster machine. Louis Monier, Paul Flaherty and Joella Paquette met over lunch. Paul wanted to test Alpha on the net and an idea emerged. Louis developed Scooter - the fastest spider ever seen, from scratch. It had to be multi-threaded and yet able to co-ordinate. The first test was run on 4 July. Digital had an indexing system developed to organise email(!). It was modified to take on the Web. The internet connection was upgraded to permit 100 Mbps. Only a company such as Digital, IBM, or equivalents, could pull this off. AV was launched on December 15 1995 and had 300,000 hits on the first day. AV claimed to be able to index the whole Web (and maybe they did for a while). It was one of the great successes in Web history. Not only were the spider and the indexer faster and better than anything before, but they also ran on HW better and faster than ever. AV also provided several front-end improvements (Chu and Rosenthal 1996):

- Natural language queries(using stop words)

- Advanced searching (incl. Boolean)
- Field search (title, URL, etc,)
- Both Web and USENET news
- Links to a particular site
- Rotating search “tips”
- Allowed users to add and delete URLs

Url: <http://www.altavista.com>

AltaVista is the premier search engine on the Web. It has the largest, most inclusive indices. That does not mean it is the only one you need, or in all situations the best one to use. Different robot and indexing strategies have resulted in different results when using the various search engines. AltaVista, however returns consistently useful information, but since no editorial decisions have been made regarding content, it also has the largest” noise to signal” ratio.

AltaVista allows searching of both the Web and many Usenet newsgroups. It allows control of the result lists in a standard, compact, and detailed format. It provides both **simple** and **advanced** searches. Advanced searches include all the features of simple ones, and allow the use of Boolean and proximity operators, grouping of terms by parentheses, and results ranking by keyword.

5.1.1 Simple searches

For an effective search, it is best to enter as many search terms or phrases which exactly qualify the subject in which you are interested as you can. The more you can be specific by offering more exact terms, the better the results.

5.1.1.1. Case sensitivity: Search terms entered in lower case letters are case sensitive. The use of the capitalised terms (or accented letters) makes the term case sensitive. *HotDog* finds only the terms spelled exactly with that capitalization: *hotdog* finds all occurrences of the term, regardless of capitalisation

5.1.1.2. Phrases: To group search terms into phrases, include them in quotation marks “*Nelson Mandela*” finds occurrences of the name *Nelson Mandela*,

capitalized in just that way. To link words into phrases is to insert punctuation between them e.g. Nelson; Mandela; Umtata: address.

5.1.1.3. Required Terms: To require that one of your terms be included in the document being indexed, preface (the formal term is prepend) it with a + symbol: + Hotdog. There must be a space between the + and the term.

5.1.1.4. Prohibited terms: To prohibit the inclusion of a term from a document for which you are searching prepend it with a – symbol: **-mustard.**

To find a reference to F, Scott Firtz without reference to Gatsby: + **“F, Scott Firtz” - Gatsby**

5.1.1.5 Wildcards: with simple queries, you are allowed to enter a wildcard character at the end of the phrases, which will substitute for any combination of letters. The asterisk (*) is AltaVista’s wildcard character. For example, butt* will get occurrences of butt, butts, butter, button. The asterisk cannot be used at the beginning or in the middle of words. It will substitute for up to 5 additional lower case letters.

5.1.1.6. Rankings: AltaVista will assign a confidence ranking to the hits it returns, based on the following:

- The query terms are found in the first few words of the document (especially the title Web pages)
- The query terms are found in close proximity to one another in the document
- The document contains more of the search terms than other documents do.

These factors are weighted, and the document with the highest confidence rating is given a score of 1.000. All others are given a decimal score less than 1.000, in order of confidence. This does not mean that the document rated 1.000 is the best source. It only meets the algorithm best. Only rarely is the “best” source ranked first, unless you know the specific title of the document for which you are searching. For example to find the document **“Mr Steven Mamatate and the internet “** a search for that phrase in double quotes will find the exact Web page, but entering the search terms separately, or just searching for **“Steven”** will result into too many non-specific hits.

Another way to search for a document with a known title is to enter the keyword title: in the search window and follow it with the title in double quotes: title: "Mr Steven Mamatate and the internet". AltaVista allows searching within specific html tags like this for anchors, applets, hosts, images, links, text, and also urls

The most useful advice for searching with AltaVista, since its indices are text based whole words, is to be as precise as possible in describing what you are looking for, while excluding things in which you are not interested. **Zimbabwe + Mugabe – conflict – war**, will find information on Zimbabwe and particularly about Mugabe, without finding information on the conflict

5.1.2. Advanced searches

The same rule for capitalisation, phrases, wildcards, required/prohibited terms, apply to advanced queries, and in addition the use of Boolean searching, proximity operators, and logical groupings with parenthesis are allowed. These are only available if you select an advanced search from the AltaVista main.

5.1.2.1. Boolean and proximity searching:

AltaVista supports the use of the binary operators AND, OR, NEAR and the unary operator NOT. You may use the following symbols in place of the words: &, ~, !, . It is a very good idea to use the words rather than the symbols, since the words are easier to remember and common to other search engines. You may enter the operators on lower or upper case letters, but it is probably best to use upper case to make them stand out from ordinary search terms and to make the logic of the search more apparent. If these words are part of the terms for which you are searching, they must be enclosed in quotation marks. It is best to group your terms within parentheses to avoid confusion, but this is not required.

Results ranking: with advanced searches you may also specify keywords you wish AltaVista to use in order to confidence rank your results. This is a very powerful feature which will let you control which items are ranked at the top of the hit list. Type the terms you wish AltaVista to weight more heavily in the results ranking

criteria box on the advanced search screen before submitting the search. Then, even though the search results will not be affected, the listing of the hits will place those in which you will probably be most interested at the top.

The searcher should proceed immediately to the AltaVista Advanced Search option in consideration of the fact that this engine indexes all existing Web pages in full text (it claims 30 million). The searcher needs every control tool offered by AltaVista to avoid being hit by a tidal wave of sites. AltaVista also offers searching of News Groups on the Web. It is not unusual for an unfiltered search to yield over 1 000 000 hits returned for a single query – in one second! One should always head straight for the advanced search mode – or for the beta page – in any search engine. This will always provide the tools for a more controlled search.

Ever since AltaVista first exploded on the scene in December 1995, it has been recognised as the premier search engine. It is regarded as being the most comprehensive of the search engines. The searcher can construct search phrases for AltaVista much like the phrases used in DIALOG and many other similar electronic databases. This has not always been the case for Internet search engines. Boolean, proximity searching, phrase searching, and field searching allowed, and can be stated in the syntax that has been well established over the years (why reinvent the wheel?). Also available are the use of wildcards (an AltaVista exclusive) and case sensitivity. Examples of “good” search engine strings for AltaVista include (Gray, 1996)

Examples

- Donkeys AND Carriages
- “Nelson Mandela” AND “criminal war”. or...(“Nelson Mandela”) AND (“criminal war”)
- (“Nelson Mandela”) AND NOT (“Criminal war”)
- “Lindiwe Mtshontshi” OR “Lennox Sebe”
- (dogs OR cats) AND (“pet care”)

Take note that the Boolean NOT must be stated AND NOT, and phrases must be placed in quotes, although the parentheses are optional. AltaVista also permits a window for the searcher to rank his search terms, a very useful device. The resulting search will be weighted to the top terms in your ranking. The user controls can help to pare down mountains of information that AltaVista is prone to provide if the controls are not used.

AltaVista has some powerful search options for experienced users: it lets you search for exact phrases, require or prohibit words, search within the title of an HTML document, search for documents that contain a link to a particular URL, use wildcards, and employ case sensitivity. The advanced search allows for the use of Boolean operators (AND, OR, AND NOT, NEAR).

Tip: Use NEAR instead of AND in the advanced search. Results will tend to be more relevant.

Example: If your question is: Which Dr. Seuss book used a vocabulary of just fifty words?, then try typing this in Advanced Search: +seuss NEAR ("fifty words" OR "50 words")

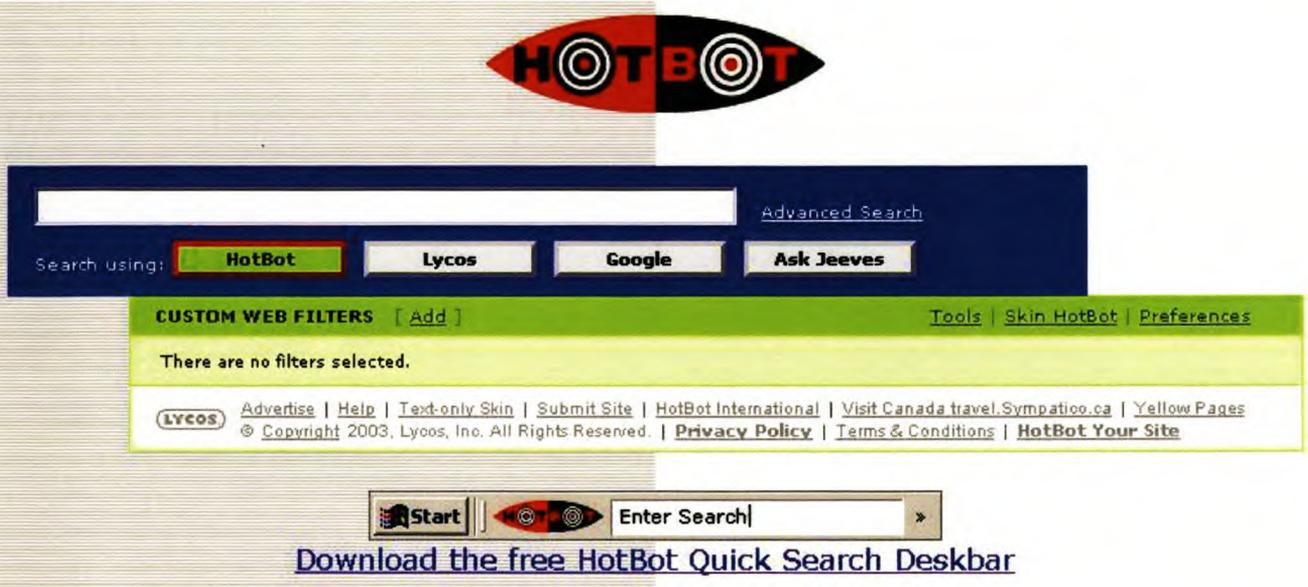
Pros: Extensive database supports complex Boolean searches. One can search in many languages. The optional Family Filter can screen out some objectionable content.

Cons: Too many hits may overwhelm novice searchers.

Help and/or FAQ Page: [Help](#), [FAQ](#)

5.2 HOTBOT

[URL:http://www.hotbot.com](http://www.hotbot.com)



It seems as if once AltaVista paved the way, Hotbot and several other search engines created Internet tools that are very similar in speed and control, but which also offer some unique features.

This search engine claims to have indexed every word in the WWW and to sit on a more flexible and powerful system than AltaVista or Yahoo. One trick to get a really specific search is to leave the first box blank. Then click on MODIFY. You can enter a word or phrase that appears. You can also click on the arrow next to the box and if you click on it will turn of.

Hotbot provides menu options to specify that a word or phrase “must” “should” or “must not “appear in the retrieved documents. The default is to search for documents containing all the search terms, but the menu allows this to be changed to “match any”, In addition, it is possible to use the operators +, - , and “ “, and also * for truncation although this last operator is not mentioned in the Hotbot help files. The search can be limited by date and by language. It is possible to limit the search to

pages containing specific items such as file extension e.g. acrobat (PDF) files. Stemming can be turned on or off. The stemming will include not only the plural form of an input search term, but other grammatical forms related to the term. It is also possible to select “Boolean Search” from the search options, which will permit the use of AND, OR and NOT and nested parentheses. A useful feature is the ability to conduct a further search on the retrieved set of documents Hotbot also provides a link to a facility to search newsgroup archives (search Usenet).

Hotbot boasts of the Boolean AND and OR phrase searching, limited by date, media type, and location in its form based menu. Once again, the experienced user should head straight for the “expert search” mode to gain maximum control of the 54 million options. A feature that permits the user limiting by media type is unique to Hotbot. With this feature, the user can access all the sites that feature specific software add-ons like JavaScript, Shockwave, Acrobat, and audio, or VRML, viewers. This is a great way to find sites that are attractive in an austere, generation-X” sort of a way. In terms of speed, all other variables considered, these major search engines are amazingly fast. Somehow, the program is able to search all 50 million sites in about one second.

HotBot supports searches using Boolean operators with nesting, and searches limited by date, location, domain names, media (image, text, sound, etc.), or page type. For searches on broad topics (about half of Hotbot searches), the top ten results come from Direct Hit, which ranks sites based on what other users with similar searches have done. The more previous searchers have clicked on a site, and the longer they've stayed there, the higher the site ranks. For more specific topics, Hotbot ranks results based on the frequency and location of your search terms.

Tip: Type the most important search terms first.

Example: If your question is: What are all the Pokemon names?

then try typing: +pokemon +names

Pros: Supports complex Boolean searching and other powerful search techniques.

Cons: The lime green colour can be considered garish. Search results often include non-relevant sites and tend to have many dead links.

5.3 EXCITE

URL: <http://www.Excite.com>



Excite is the first engine discussed here that qualifies as both an effective Web directory organised by category and a Web search engine. It also lists millions of indexed URLs, so it cannot be criticised for having a smaller pool of pages than other Web directories. The user can search the text of at least 10,000 newsgroups, a daily news summary, opinion columns, cartoons, and Web site reviews.

Excite uses Intelligent Concept Extraction to find relationships between words and concepts. Retrieved documents will contain not only the input search items, but also words that are conceptually related to these terms. The relationships are built up during Exile's indexing procedure.

In the normal search mode, Excite accepts the following: “ - ”, +, -, AND, OR, AND NOT and permits the use of parentheses. The “power search” does not permit any of these operators, but instead offers the same facilities through a menu system. This can be used to specify that the retrieved documents “must”, “can” or “must not” contain a specified search word or phrase. If one of the retrieved documents is of particular relevance, it is possible to request “more like this”. The index terms assigned by Excite to this document are then used as the basis for a further search.

Excite allows searching Keywords or concepts and offers searching in all the areas mentioned below: UseNet newsgroups, reviews, Web documents, or classifieds. Allowable Excite search terms include (Gray, 1996):

- (illegal AND immigration) AND NOT (California)
- alien OR UFO

- alien AND NOT UFO
- Football AND (rugby OR soccer)

It also offers an option to retrieve “more like this”, a kind of citation pearl-growing feature (“query by example”) as Excite calls it, which is an essential ingredient in so many sophisticated electronic databases today. The user can pick a document that is a good match for the desired reference question, click a button next to it, and automatically reinitiate the search using the indexed search criteria of this document. This useful feature seems to be unique to Excite. The fact that Excite is not only a search engine but also a Web directory provides it with the information to make these “see also” type recommendations.

However, the tests that I performed on Excite included trying to access my name on the home pages using my own specified search terms. They produced some very strange results. The Excite screen is clustered when the researcher keyed in “Lindiwe Mtshontshi” and got 2 hits, then keyed in “MTSHONTSHI” and got 30 hits - this time I was hit number 7. This is very inexplicable to me, so I will not attempt it here. Suffice to say, if I cannot get predictable results when I key in my own search terms, I tend to generally distrust the keyword matching ability of this engine across the board.

The Excite screen is cluttered and more than a little obtuse. Do not bother clicking on the advanced search link unless all you are after is information, because you cannot enter search terms from the advanced screen, you have to back out to the original screen to perform a search.

5.4. LYCOS

URL:<http://lycos.com>



Many people who have used the Internet for a while have a fond spot for Lycos. Since the explosion of the Web, better search engines have appeared, but Lycos is still good and fast, if not as sophisticated as some of the others.

It offers both keyword and subject searching (the subject searches are called directory services), as well as a point rating system which rates Web pages. Its strong points are its speed, ease of use, and the large size of its indices, which often produce usable results by sheer force. Its weakest point is that it does not support Boolean searching or any of the more sophisticated searches that can be made with AltaVista, WebCrawler or Excite.

At any rate Lycos is still quite popular, but objectively speaking, it hasn't quite kept pace with some of the newer shinier engines. Lycos allows the internet user to:

1. "Search for specific subject or destinations
2. browse interesting categories
3. (and) have a guided tour through sites of interest" (Lycos , 2003)

5.5 WEBCRAWLER

URL: <http://www.WebCrawler.com>

The screenshot shows the WebCrawler search engine interface. At the top, there is a logo for WebCrawler with a spider icon. Below the logo, there are three tabs: "Web Search", "Yellow Pages", and "White Pages". Under "Web Search", there are three radio buttons for "Web Pages", "Photos", and "News". A search input field is present, followed by a "Search" button. Below the input field, there is a checkbox for "Exact Phrase" and links for "Advanced Search", "Preferences", and "Tools & Tips". The main content area is divided into three sections. The left section promotes WebCrawler's ease of use and lists partner engines: Google, Yahoo, Ask Jeeves, About, LookSmart, Overture, Teoma, and FindWhat. The middle section features a "Speechless?" graphic and promotes the "WebCrawler Toolbar Thesaurus". The right section lists "FEATURED SEARCHES" including IRS, Alex Rodriguez, Academy Awards, Prom Dresses, Dr. Phil, and Used Cars. At the bottom, there are links for "Set Your Homepage!", "Submit Your Site", "Terms of Use", "Privacy Policy", "Contact Us", and "Popular Searches". A copyright notice at the very bottom reads "© 2004 InfoSpace INC. All Rights Reserved".

WebCrawler was begun in 1994 at the department of Computer Science and engineering at the University of Washington. It was the very first search engine on the net. WebCrawler was purchased by America online in 1995 and became part of Excite Corp. in 1997. The WebCrawler indexes are built both by user submission and by “Spidey” the Web crawler, featured on the very inviting search interfaces. WebCrawler searches employ artificial intelligence technologies from Personal Library Software (PLS). Boolean operators, nested logic, proximity operators and bound phrases are all used in search queries, but it is not possible to disengage or override the PLS fuzzy logic capability which purports to allow natural language searching. WebCrawler displays maps interactively when the logic detects a geographic reference. A formless search option is available. There is extensive online help for general searching and advanced search techniques. WebCrawler allows users with high-end browsers that support cookies to set search preferences.

As previously mentioned, WebCrawler has the smallest index of the major search engines, estimated at 500,000 URL’s (Sullivan, 2002). It does index its sites full text,

but WebCrawler's principal criteria for selecting sites to add to the index is page popularity, or the sites that are the most well travelled in terms of visitors. To my mind, this method would tend to yield sites that are "pop" in nature, or concerned with mainstream information.

Another problem is that only the page titles of each retrieved URL are displayed for the searcher. This title may or may not be descriptive enough to provide intellectual access to the documents. The searcher is forced to link each page to sense what the content is about.

If the object of your search is mainstream information, such as information in high-profile corporations, television networks, sports, or movie stars, WebCrawler should be your first choice. This is more the character of this index, and it does occupy a distinct niche. I must add however, that judicious use of controlled languages when using the more comprehensive engines like AltaVista, Hotbot, or Info seek ultra, should enable the searcher to locate the same material.

WebCrawler is fast and easy to use. It does offer a browsable subject catalog, and in the "advanced mode" it offers Boolean and proximity searching to hone your search, but once again, WebCrawler's index is only 1% of the size of the big indexes, so I really cannot conceive of a good reason for using it as a search engine," Compared with the newer speed merchants such as AltaVista and Hotbot, WebCrawler isn't the fastest or most up to-date search engine".

Excite has noted that WebCrawler is designed to be simple and easy to use. But don't let the whimsical user interface fool you. Behind that cute little surfing spider lies one of the most powerful search technologies on the net.

5.6 YAHOO

URL:<http://www.yahoo.com>



Yahoo is not a search engine, but strictly, a hierarchically arranged subject index. It has developed over a long time, with lots of editorial care, so the quality is very high. Browsing Yahoo is the best way to surf for good sites when you don't know (or perhaps care) where exactly you are going. It is also the best way to find a good starter site, from which you can branch out to more specialized ones (Yahoo, 2003).

Using Yahoo is simple: just enter your search term(s) in the search window and click SEARCH. Yahoo will return the following

- 1) Yahoo categories that match the search (so you can explore them for cross-referencing;
- 2) Actual matching end-sites; and
- 3) The Yahoo categories from which the various pages are indexed – sort of a 'much broader term' cross reference. Though you cannot create very sophisticated search statements, you can control:

- where to search : Yahoo (default) ,Usenet or Email Addresses
- whether to OR or AND (default)the search terms
- Whether to search on substrings (find whole word from partial strings— like headlines when searching for head) or complete words (find headlines only when entering the term headlines). Substrings are the default.
- Control the number of matches per page to 10,25 (default), 50 or 100

You may access these controls by clicking the small 'option' link next to the main search window.

Yahoo has a couple of other unique features: At the bottom of each results page links to search engines are provided. By clicking on Yahoo Remote, you can invoke a secondary Netscape window, which you can minimize and then maximize whenever you need to do a quick search.

If the essential search engine is AltaVista, the essential subject catalog is Yahoo! Don't surf without it.

Database: A subject guide and search service of Internet resource news, maps, classified advertising, stock quote, sports scores, businesses, telephone numbers, personal Web pages, and email addresses (separate databases).

Content: Main directory: Links (URLs) to Internet resources and brief descriptive text for those links.

Searching: All Yahoo pages include a simple search box, but the Yahoo Search page offers options for searching Yahoo! Usenet, or Email Addresses. Searches can also be limited to listings added in the last day, week, month, or three years. Boolean operators (AND, OR) and string searching are also supported. Note: if no results are retrieved in Yahoo, the search is automatically passed to Alta Vista, which then searches its database and passes the results back via Yahoo. If Yahoo cannot make the connection fast enough with Alta Vista, it will provide a page of links to a number of search tools. When one of those links is selected, your search words are automatically passed to the search engine on your behalf. (Schilichting 1999).

SEARCH TIP: Yahoo! is a subject directory, which means it will not list many pages that search engines will typically retrieve (such as Joe Schmoe's page of hot links). Use a few words that describe your topic or that may be found on a high-level page (the first page you would see when visiting a site) for an organization or company.

Results: **links** are returned along with their descriptive text, and the subject hierarchy under which it can be found in Yahoo is also displayed. The search term(s) appear in bold on the results screen.

Address: <http://www.yahoo.com/>

Update frequency: daily

Additional Information: More information can be found at the Yahoo information Centre, including Yahoo Help.

Contact: See the Writing to Yahoo! page or send email to Webmaster@yahoo.com

5.7 INFOSEEK

This search engine was introduced in August 14 1996, and offers a major improvement over its predecessor, Infoseek guide, which is still very much alive. This is a very impressive product that also boasts of having over 50 million URLs in its index, but what really sets it apart from the others is what Infoseek calls its “real time index” of the Web (Grady, 1996). This rather obtuse phrase really means that Infoseek is actually updating its index continuously. Its spider senses new and changed pages and updates the index immediately.

Infoseek was once the only Netscape default search engine. It is not the best available. Its virtues are speed and ease of use. Its defects are a lack of sophistication (Booleans are not supported) and a teaser approach to showing the first 100 hits and offering to show more for pay. It is a search engine, and offers searchable subject catalogs, with options to search Usenet newsgroups, e-mail addresses and Web FAQs.

Searches are quasi-case sensitive. Capitalized words are taken as proper nouns and the search is limited. Searching for ‘Babe’ will find the famous hitter and the famous pig, searching for “babe” also finds the Sonny and Cher lyrics. Adjacent capitalized words link them into phrases. Capitalized phrases must be separated with commas e.g. the great bambino, baseball hall of fame. Phrases may be formed by enclosing the words in double quotes e.g. “I’ve got you babe”, yet a third way to link words into phrases is to place hyphens between them: wonderful-life.

Required/prohibition operators. Prepending a word with a + symbol it requires that the term must be in the documents found by the search. Prepending a – symbol excludes documents containing that term from the search results: Mandela-automobile. There may not be a space between the + or – sign and the affected word.

Proximity operator: Placing words in square brackets causes a hit if they are found 1000 words from each other

Willis advises that to search Infoseek 'select sites' (their subject catalog) change the search option from World Wide Web to "Infoseek Select Site" on the form provided next to the search term window. There are other several options available.

Some estimates claim that almost half of the URLs on the Web are either duplicates or dead/invalid links (Infoseek, 2002). Infoseek ultra has created software that filters out duplicates and/or dead links, and this is a major feature of this engine. The researcher is yet to get an invalid link message in any of her infoseek ultra searches. These searches are lean and accurate, with a very high "signal-noise ratio" also known as high precision.

5.8 OPENTEXT

URL: <http://www.OpenText.com>



Open text has been in a state of flux from its early days, so the information in the help pages, if you can find them, is no longer accurate. Features and navigation have changed. It is still, however, an excellent search tool.

The default search window is what used to be called the Power Search. Basically, it represents 3 searches into which a word or phrases can be entered, separated by a qualifier as to where to search (anywhere (default), document summary, title, first heading, or URL) and also separated by Booleans (AND (default), OR +, BUT NOT, NEAR). The instructions are as follows:

Search for {enter your search term(s) within {choose where} {Boolean option to connect to next search term}. Three terms can be entered and qualified.

Opentext does not support a wildcard expansion character, but does handle plurals nicely. Do not enter plural search terms. Opentext will search for plurals automatically, including plurals like “geese”.

Booleans: The attempt to make use of Booleans and proximity operators simple has backfired. Entering the actual operators and grouping terms with parentheses is far easier and quicker than selecting from boxes. Understanding the logical interpretation of the operator is also more difficult when laid out in linear fashion like this.

Proximity Operators: Opentext implements both NEAR operators, with a non-adjustable range of 80 words, and is followed by an operator like WebCrawler's ADJ operator where word order matters – once again, with a non-adjustable range of 80 words. Such a large range reduces the usefulness of these operators.

Opentext does not limit whole words, so that a search for the word “head” will also get hits on “headstrong” and “headline”. It will also miss terms if entered in plural rather than singular. Exact, correct spelling is important with Opentext.

Very good features include the ability to see the terms from the referenced pages that caused the hit.

CHAPTER 6

6.1 DISCUSSION AND CONCLUSION

Before a researcher logs onto the Internet, he needs to answer a few simple questions to help him determine the best type of search tool for his purpose. If he is looking for specific information, the best choice is a search engine in the order of preference. If the purpose is merely to browse sites to learn what is available on the subject of interest, the subject indexes are the place to start. The meta-search engines are alluring, but theoretically at least, search engines that are comprehensive like AltaVista, Hotbot, Infoseek ultra, Lycos and Excite should yield much the same results. When using these comprehensive engines, the searcher needs to be as explicit as necessary to retrieve the level of results desired. Also, if precise information is needed, the search terms likewise need to be as precise and limiting as possible. As previously mentioned, AltaVista seems to be the best at matching the level of search terms with its level of retrieved documents, and for this and many other reasons, is my first choice for an internet search tool.

Search engines, if used properly are able to match search terms with corresponding terms contained in specific Web sites. Many of the newer engines incorporate a spider or robot software to index Web sites. This automated process actually visits each new Web page and records the full text of every page (including as many as three of page links). Other engines may only base their indexing on the title, heading, and say the first 200 words of the body. Still others may analyze the number of links that point to the page being indexed, to determine its usefulness. The point is, each search engine goes about the job of indexing in a different way. The other half of the process, the front end offered to the user via the search screen, also varies widely in terms of the operations and features engineered into the software. Some engines permit the user to key in all the necessary control language elements such as Boolean operators, and various limiting schemes. Others simply present forms and pull down menus that allow the user to select the proper limiting terms. The later technique is referred to as “form based” controls. The bottom line is that search engines rarely yield identical results when presented with identical search terms. The user, if she wishes to use each

engine effectively, needs to understand the differences in the construction and the use of each, in order to make an informed choice of product.

All search engines match the user's search terms to documents in roughly the same way (Sullivan 1996). These are simply:

- Keywords are in the first few words of the document (keywords in title subtitle etc.)
- Keywords are found close to one another in a document (keyword proximity)
- Documents contain more of the query words than others (keyword frequency counts).

If this all sounds strangely familiar to DIALOG, OPAC, and electronic database searching in general, it should. The concept isn't essentially different. However, the concepts have been transposed and rechristened by many of the familiar search engines – much to the consternation of those of us who understand the principle involved. The best of the search engines, AltaVista, HotBot, Info seek and Excite do offer the searchers well-established controls that are critical for weeding the millions of sites that exist on the Web.

Subject catalogs are actually hierarchically organised indexes of subject categories that permit the searcher to browse through lists of Web sites by subject in search of relevant information (Tyner, 1996). The analysis of sites by subject is done by humans, not computers, and therein lies both their advantage and disadvantage. First the advantage: the pool of indexed sites is necessarily smaller in comparison to search engines that use an automated robot spider to collect indexing information. However, no amount of word frequency counting or proximity calculation can compare with the interpretative ability of the human mind. So, when browsing a subject catalog, one can be assured of subject relevancy (high precision), but not comprehensiveness (high recall). This is the best answer for the poor researcher!?!

In the case of search engines, the more powerful the controls the searcher has to sort and manipulate the hits in a predictable and intuitive fashion, the better. As in all other forms of electronic querying, the user simply must take time beforehand to

analyze and list as many relevant synonyms and necessary terms as possible the more precise the query, the more likely that material retrieved will be useful.

The searcher also needs to consider the level of responses needed. To state this concept simply, the user may want to approach the subject very broadly in order to gain ideas of just how large the body of information is that is relevant to his topic. On the other hand, he may want to be very specific, exacting information about the topic to answer questions or help to confirm a hypothesis.

There are many, many search engines available with comparable capabilities. Most search engines now have basic and advanced search features. Basic searches simply allow you to enter a search in whatever form or format you choose. Advanced search features allow you to give the search engine more specific instructions for executing your search, and give you more control over the structure of your search. Once you have explored the search capabilities of a particular search engine, then you can choose the best one(s) for you. Broad or general terms will return thousands of possible sites. Try to use terms that are more specific to your topic. To narrow your terms, look at sites that you already have found and that are relevant to your topic. Identify possible search terms from those sites. You also can combine terms using Boolean Operators

6.2 RECOMMENDATIONS - *Choosing the search engine*

“When love and skills work together, expect masterpieces” John Ruskin. Have you ever had a hard time finding the information that you want on the Web? You are not alone. The best way to defend yourself from too much non-relevant information is to prepare a good defence. Arm yourself with Boolean logic, and strengthen your search strategy by understanding the difference between directories and search engines. When one search engine isn’t getting you anywhere, you can always try something else Internet search engines are constantly changing, with new ones appearing and new features being added to existing ones. This section does not attempt to describe all currently available search engines, but provides a comparison of some, which are

deemed to be among the most useful. The choice of search engines will be dictated by a number of factors. There are differences in the number of Web sites covered. The depth to which they are covered (i.e. whether all, some or just the home page is indexed), and the frequency with which sites are re-indexed. For example data from Searchenginewatch, a site which compares a number of different search engines, gives the following figures for the number of pages indexed in millions as of January 2002, AltaVista 500, Northern Light 208, Excite 225, Lycos 500. These figures need to be compared to an estimate made in "Nature" in July 2001 that approximately 800 million pages were available on the Internet: a figure, which will have increased since then. The frequency of re-indexing varies considerably; with AltaVista and Info seek showing the widest ranges of 1 day-1 month and 1 day-2 months, respectively, Info seek and WebCrawler only sample a proportion of pages from any Website while the other search engines claim to index all these pages within a site eventually. The overall speed with which a search engine can be accessed, and the speed with which it processes a query, may be important in determining choice. For example, Lycos Pro offers some very sophisticated features, but frequently downloads so slowly as to be almost unusable. Speed is one reason why some of the smaller and probably less accessed regional search engines may also be useful.

The available search features vary in sophistication and are compared below. The orders in which the search results are displayed will depend on the algorithms used by these engines to process the results and assess relevancy. These change frequently so the same search may give different results from week to week. The criteria used by the search engines to determine relevance may not necessarily relate to the relevance of the site content. Looking at some of the lower rankings may sometimes pull up highly relevant pages. Another strategy is to carry out a search on a meta-searcher. These are not search engines as such, but they process queries for simultaneous submission to a number of different search engines, so that pages are identified based on a number of sets relevancy criteria. It should always be remembered that many relevant pages on the Web may not be identified as such by search engines. One of the best ways of finding these pages is by following links supplied by other relevant pages, which have been retrieved. A number of meta-sites consist only of links to pages concerned with one or more related topics. When following a large number of links, the best strategy is to right click on each link in turn and then select to either bookmark it for a future evaluation folder, or open it in a new browser window. Once

the link has been evaluated and bookmarked if useful, the window can be closed, leaving visible the first window with the original list of links Lynch (1997) states that it is quite simple; the choice of search engine is entirely dependent on the needs and preferences of the searcher. These needs can be every bit as diverse as the internet itself. Taking a very broad overview, search engines are the tool of choice when the searcher has a specific question in mind. They are prone to deliver very high recall, so it is imperative that they offer features that allow the searcher to narrow and limit the search. On the other hand, the subject catalogues are more appropriate for browsing the Net, and their retrieval characteristics can be described as high precision.

Future research is needed to further test and evaluate the value of the measures proposed. The strength of the IR evaluation is based on the strength of the models that underpin its development. Further research is needed that looks beyond traditional approaches to IR evaluation to consider the information-seeking context of the user.

6.3. RECOMMENDED SEARCH ENGINE

Some ways the recommended search engines differ, Yahoo is the best engine among these seven because of the following features:-

Links to help – Yahoo help pages

Size, type – Huge over 3 billion fully indexed, searchable pages

Noteworthy features and limitations – No 10 word limit

Results ranking – Relevancy ranking

Stemming – None

Foreign accents – None

Boolean logic – Full accepts AND, OR, NOT or AND NOT () for nesting Must be capitalized

Field limiting - link; site; intitle; inurl; hostname; linkdomain;

Limit by age of documents – In advanced search

Translation – yes

Shortcuts – shortcut give quick access to Calculator, dictionary, synonyms, patents, traffic, stocks, encyclopedia, and more

Specialized databases – Directory, Images, News, Products, Yellow pages, Groups and more

REFERENCES

- Andrew, W (1996) Search engines gain tools for sifting content on the fly. *Web Week*, 2(11), 41-42
- AltaVista Search: questions (1999) Available:
www.altavista.digital.com/av/content/ques_master.htm
- Bar-Ilan, J (2002) Methods for measuring search engines performance over time: *Journal of the American Society for Information Science and Technology*, 53(4), 308-319
- Barry, C.L. and Schamber, L. (1998) User criteria for relevance evaluation: A cross-situational comparison. *Information proceeding and Management*, 34(2/3), 219-236
- Bates, M.J. and Schamber, L (1998) User's criteria for relevance evaluation: A cross-situational comparison. *Information Processing and management*, 34(2/3) , 219-236.
- 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: http://asis.org/annual-96Electronic_proceedings/chu.html
- Clarke, S.J. and Willet, P. (1997) Estimating the recall performance of Web Search Engines. *AslibProceedings*, 49(7), 184-189.
- Cleverdon, C.W. (1978) User evaluation of information retrieval systems, In: King, D (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 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 ordering action of retrieval system, *Journal of the American Society of Information Science*, 19(1), 30-40

- Cooper, W.S. (1973) On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24, 87-100
- Courtois, W.S. and Berry, M.W. (1999) Results ranking in Web search engines. *Online*, 23(3)
Available:<http://www.online.com/onlinemag/OLtocs/OLtocmay4.html>
- Croft, W.B. (1995). What do people want from information retrieval? *D-Lib Magazine*, November Available :
<http://www.dlib.org/dlib/novmenr95/11croft.html>
- DeMoss, Timothy, (1996) Gentlemen, Start your Engine , *Power engineering*, August 1996 p10-13
- 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
- Feldman, S (1998) Web Search services in 1998:trends and challenges. *Searcher*, 16(6) Available: <http://www.infortoday.com/it/jun/felman.htm>
- Feldman, S. (1999) Search engines: the 1999 conference. *Information Today*, 16(6).
Available: <http://www.infortoday.com/it/jun/felma/htm>
- Gatian, Army W. (1994) Is user satisfaction a valid measure of system effectiveness? *Information and Management*, 26, 119-131
- Goodhue, Dale, L (1995) Understanding user evaluation of information systems. *Management Science* 41(12), 187-1843
- Gordon, M., & Pathack, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing and management*, 35(2), 141-180
- Grady, Steve, (1996) Infoseek introduces Infoseek Ultra, Hypertext document
Available:
<http://C%7C/Program%20Files/Netscape/Na...32.19960930112925.0006d5d78@corp&numbe=51>

- Harter, Stephen, P. and Hert, Carol, A. (1997) Evaluation of Information retrieval system: approaches, issues, and methods. In: Martha E. William (ed). *Annual Review of Information Science and Technology (ARIST)*, Vol 32, 3-94
- Jansen, B.J. and Spink, A. (2000). Methodological approach in discovering user search patterns through Web log analysis. *Bulletin of the American Society for Information Science*, 27(1). 15-17
- Jansen, B. J. & Pooch, U. (2001) A review of Web searching and a framework for future research . *Journal of the American Society for Information Science and Technology*, 53(3) , 235-246
- Koboyashi, M. & Tekadi, K (2000) Information retrieval on the Web. *AMC Computing Surveys*, 32(2), 144-173.
- Koll, M. (1993) Automatic relevance ranking. A searcher's 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) *Information Retrieval systems: characteristics, testing and evaluation*. 2nd edition New York: John Wiley
- 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 Available: <http://www.dlib.org/april97/04larsen.html>
- Leighton, H.V (1996) Performance of four World Wide Web Index services: Infoseek, Lycos, WebCrawler and WWW Worn. Available: <http://www.winona.msus.edu/is-f/library-f/Webind.htm>
- Leighton, H.V. (1997) Precision among World Wide Web search services search engines Alta Vista, Excite, HotBot, Info seek, Lycos. Available: www.winona.edu/library/Webind2/Webind2/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), 970-881

- Lui, Jian, (1996) Understanding WWW search tools Reference libraries, IVB Libraries. Available: <http://www.indiana.edu/~libresd/search>.
- Lui, Jian, "Understanding WWW search Tools", Reference Department, IVB Libraries. Available: http://www.library.ucsb.edu/isti/98-spring_article5.html
- Lynch, Clifford (1997). Searching the Internet. *Scientific American* March 1997, pp52-56
- McDonald, Jason, (1996) Simple route to more efficient Web searcher *Machine design* 24 October 1996 pp78-80
- Mizzaro, S. (1997) Relevance: The whole history. *Journal of the American Society for Information Science*, 48, 810-832
- Notess, Greg. (2000) Search engines shutdown Available: <http://www.notess.com>
- Oppenheim, C. Morris, McKnight, A. and Lowly 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
- Richard, Peter and Robert , Sikorski, (1997) Smarter Searching, *Science* 27(2) August 1997, pp976-979
- Salton, G. (1992) The state of retrieval system evaluation. *Information processing management*, 28(4), 441-449
- Salton, G. and Mc Gill, M. J. (1983) *Introduction to modern information retrieval*. McGraw-Hill: New York
- Sandore, B (1990) Online searching: what measure satisfaction? *Library and Information Science research*, 12, 33-54
- Schlichting, C., & Nilsen, E (1996). Signal detection analysis of WW search engines. Presented at Microsoft's designing for the Web: *Empirical studies* Conference, October 1996 Retrieved December 2003 from <http://www.microsoft.com/usability/Webconf/schilichting/schilichting.htm>

- Schilichting, Carton & Nilsen, Erik.(2000) Signal Detection Analysis of WWW Search engine” Available:
<http://www.microsoft.com/usability/Webconf/Schilitching.htm>
- Silverstein, C. Henzinger, M. Marais, J & Moric, M (1998) Analysis of a very large AltaVista query log. *Technical Report* 1998-014, COMPAQ Systems Research Center, Palo Alto, Ca, USA 1998
- Spin, A, Dietman, (et al) (2001) Searching the Web: the public and the queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-324
- Spink, A., Wilson, T., Ellis, D., and Ford, N. (1998). Modeling users successive searches in digital environments. *D-Lib Magazine*, April 1998 Available:
<http://www.dlib.org/dlib/april98/04spink.html>
- Su, L (1992) Evaluation measures for interactive information retrieval. *Information Processing and Management* 28(4). 503-516
- Su, L and Chen, H. (1999) User evaluation of Web search engines as prototype digital library retrieval tools. *Information Processing and Management* 28(2) 407-500
- Sullivan, D (1996). Survey reveals search habits. *Search Engines Report* Available:
<http://www.searchenginewatch.com/sereport/00.06-realnames.html>
- Sullivan, D. (1998) Counting clicks and looking at links. *Search engines report* Available: <http://www.searchengineswatch.com>
- Sullivan, D. (2000) Web search engine trends and achievements since the 1999 Boston Search Engines meeting, In *Search Engines Today and the New Frontier: The fifth Search Engine Meeting* Boston, Massachusetts, April 2000. Available: www.infonortics.com/searchengines/boston/boston2000pro.html
- Sullivan, D. (2002) Web search engine trends and achievements since the 1999 Boston search engine meeting. In: *Search Engine today and the New Frontier: The fifth Search engine meeting in Boston, Massachusetts, April 2002* Available: <http://www.infonortics.com/searchengines/boston2002pro.html>
- Travis, I.. (1998) From storage and retrieval systems to search engines: text retrieval in evolution. *ASIS Bulletin*, April/May, 2p

- Tyner, Ross (1996) Sink or swim: Internet search tools and techniques. Okanagan University College. Hypertext document: Available:
<http://www.sci.ouc.bc.ca/libr/connect96/search.htm>
- Venditto, G. (1996). Search engine showdown: IW labs test seven Internet search tools. *Internet world*, May, 79-86
- Westera, Gillian, (1996) "Search engine Comparison: Testing retrieval and Accuracy"
Available: [http://www.asis.org/annual-96/electronicc Proceedings/chu.html](http://www.asis.org/annual-96/electronicc%20Proceedings/chu.html)
- Wiggins, R. and Matthews, J (1998). Plateaus. Peaks and promises: the Infonotics'98 search engines conference. *Searcher*, 6(6) online Available:
<http://www.infortoday.com/searchers/jun98/story4.htm>
- Wiley, D. (1998) Beyond information retrieval *Database* 21 (4) Available:
<http://onlineinc.com/database/DB1998/wiley8>
- Willis, Lynn. (1996) Touring the Internet *Chemtech* July 1996, pp19-20
- Wishard, L (1998) Precision among Internet search engines: An earth sciences case study issues in Science and technology Librarianship 1998 Available:
<http://www.library.com>
- Wiship, I.R. (1995) World Wide Web searching tools –an evaluation. *VINE*, 99, 49-54
Available: <http://bubl.bath.ac.uk/BULB/winship.html>
- Zorn et al (1999).Data mining meets the Web. Online September/October, 17-28
Available: <http://www.onlineinc.com/onlinemag/>