Evaluation Of Internet Search Tools
Instrument Design

Tana Saunders

Assignment submitted in partial fulfillment of the requirements for the degree of Master of Philosophy (Information and Knowledge Management) at the University of Stellenbosch.

Supervisor: Dr Martin van der Walt

April 2004

DECLARATION

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

ABSTRACT

This study investigated Internet search tools / engines to identify desirable features that can be used as a benchmark or standard to evaluate web search engines. In the past, the Internet was thought of as a big spider's web, ultimately connecting all the bits of information. It has now become clear that this is not the case, and that the bow tie analogy is more accurate. This analogy suggests that there is a central core of well-connected pages, with links IN and OUT to other pages, tendrils and orphan pages. This emphasizes the importance of selecting a search tool that is well connected and linked to the central core. Searchers must take into account that not all search tools search the Invisible Web and this will reflect on the search tool selected. Not all information found on the Web and Internet is reliable, current and accurate, and Web information must be evaluated in terms of authority, currency, bias, purpose of the Web site, etc. Different kinds of search tools are available on the Internet, such as search engines, directories, library gateways, portals, intelligent agents, etc. These search tools were studied and explored. A new categorization for online search tools consisting of Intelligent Agents, Search Engines, Directories and Portals / Hubs is suggested. This categorization distinguishes the major differences between the 21 kinds of search tools studied. Search tools / engines consist of spiders, crawlers, robots, indexes and search tool software. These search tools can be further distinguished by their scope, internal or external searches and whether they search Web pages or Web sites. Most search tools operate within a relationship with other search tools, and they often share results, spiders and databases. This relationship is very dynamic. The major international search engines have identifiable search features. The features of Google, Yahoo, Lycos and Excite were studied in detail. Search engines search for information in different ways, and present their results differently. These characteristics are critical to the Recall/Precision ratio. A well-planned search strategy will improve the Precision/Recall ratio and consider the web-user capabilities and needs. Internet search tools/engines is not a panacea for all information needs, and have pros and cons. The Internet search tool evaluation instrument was developed based on desirable features of the major search tools, and is considered a benchmark or standard for Internet search tools. This instrument, applied to three South African search tools, provided insight into the capabilities of the local search tools compared to the benchmark suggested in this study. The study concludes that the local search engines compare favorably with the major ones, but not enough so to use them exclusively. Further research into this aspect is needed. Intelligent agents are likely to become more popular, but the only certainty in the future of Internet search tools is change, change, and change.

OPSOMMING

Hierdie studie het Internetsoekinstrumente/-enjins ondersoek met die doel om gewenste eienskappe te identifiseer wat as 'n standaard kan dien om soekenjins te evalueer. In die verlede is die Internet gesien as 'n groot spinnerak, wat uiteindelik al die inligtingsdeeltjies verbind. Dit het egter nou duidelik geword dat dit glad nie die geval is nie, en dat die strikdas analogie meer akkuraat is. Hierdie analogie stel voor dat daar 'n sentrale kern van goed gekonnekteerde bladsye is, met skakels IN en UIT na ander bladsye, tentakels en weesbladsye. Dit beklemtoon die belangrikheid om die regte soekinstrument te kies, naamlik een wat goed gekonnekteer is, en geskakel is met die sentrale kern van dokumente. Soekers moet in gedagte hou dat nie alle soekenjins in die Onsigbare Web soek nie, en dit behoort weerspieël te word in die keuse van die soekinstrument. Nie alle inligting wat op die Web en Internet gevind word is betroubaar, op datum en akkuraat nie, en Web-inligting moet geëvalueer word in terme van outoriteit, tydigheid, vooroordeel, doel van die Webruimte, ens. Verskillende soorte soekinstrumente is op die Internet beskikbaar, soos soekenjins, gidse, biblioteekpoorte, portale, intelligente agente, ens. Hierdie soekinstrumente is bestudeer en verken. 'n Nuwe kategorisering vir aanlyn soekinstrumente bestaande uit Intelligente Agente, Soekinstrumente, Gidse en Portale/Middelpunte word voorgestel. Hierdie kategorisering onderskei die hoofverskille tussen die 21 soorte soekinstrumente wat bestudeer is. Soekinstrumente/-enjins bestaan uit spinnekoppe, kruipers, robotte, indekse en soekinstrument sagteware. Hierdie soekinstrumente kan verder onderskei word deur hulle omvang, interne of eksterne soektogte en of hulle op Webbladsye of Webruimtes soek. Die meeste soekinstrumente werk in verhouding met ander soekinstrumente, en hulle deel dikwels resultate, spinnekoppe en databasisse. Hierdie verhouding is baie dinamies. Die hoof internasionale soekenjins het soekeienskappe wat identifiseerbaar is. Die eienskappe van Google, Yahoo en Excite is in besonderhede bestudeer. Soekenjins soek op verskillende maniere na inligting, en lê hulle resultate verskillend voor. Hierdie karaktereienskappe is krities vir die Herwinning/Presisie verhouding. 'n Goedbeplande soekstrategie sal die Herwinning/Presisie verhouding verbeter. Internet soekinstrumente/-enjins is nie die wondermiddel vir alle inligtingsbehoeftes nie, en het voor- en nadele. Die Internet soekinstrument evalueringsmeganisme se ontwikkeling is gebaseer op gewenste eienskappe van die hoof soekinstrumente, en word beskou as 'n standaard vir Internet soekinstrumente. Hierdie instrument, toegepas op drie Suid-Afrikaanse soekenjins, het insae verskaf in die doeltreffendheid van die plaaslike soekinstrumente soos vergelyk met die standaard wat in hierdie studie voorgestel word. In die studie word tot die slotsom gekom dat die plaaslike soekenjins gunstig vergelyk met die hoof soekenjins, maar nie genoegsaam so dat hulle eksklusief gebruik kan word nie. Verdere navorsing oor hierdie aspek is nodig. Intelligente Agente sal waarskynlik meer gewild word, maar die enigste sekerheid vir die toekoms van Internet soekinstrumente is verandering, verandering en nogmaals verandering.

ACKNOWLEDGEMENTS

I wish to acknowledge the indispensable support and encouragement
received during the past two years from
the following significant people in my life:

My beloved and courageous husband Nick, and Chocolate, the Siamese
My loving parents
My longsuffering friends
And my inspiring study leader.

Thank you for your encouragement, suggestions, ideas and unfailing love.

Most of all, I thank Our Heavenly Father, who leveled the mountains…

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 BACKGROUND

There is a great deal of useful information available on the Internet but not even the most resilient searcher could follow hyperlinks to all the documents on the World Wide Web and Internet. There are millions of pages and billions of words, in varying formats and languages (Barlow, 2002b). Every minute of every day, more documents are posted, more information is made available – the phenomenon of Information Overload personified (Hölscher & Strube, 2000: 337). To access the sea of information, 85% (Savoy & Picard, 2001: 543) of searchers use Internet Search tools, generally referred to as search engines, and users are eager to know how these search engines compare (Bharat & Broder, 1998:379).

1.1 Problem statement

Not all Internet search engines are equally efficient and effective in retrieving reliable, valid and accurate information. The fact that there are a growing number of search engines available complicates the decision of which one or which combination of search engines to use in an Internet search.

This study explores the problem of Web search engine / search tool evaluation. The study focuses on developing a methodology or instrument that will be useful to evaluate Internet search tools / engines to optimize the Precision/Recall ratio and improve user satisfaction.

The target group for the evaluation instrument / tool are both searchers and users – professional searchers for Web based information, as well as the occasional, non-professional Internet user that requires reliable and current information.

Desirable search engine features and functionalities are identified by studying different search engines and by addressing the research questions in the table below. The answers to these research questions informed the development of the evaluation tool/instrument.

	Research questions	Motive and relevance to the evaluation instrument
1	Is information found on the WWW reliable, and how can it be evaluated?	The information retrieved by the search engine must be evaluated to gauge relevancy to the search query; general information evaluation criteria apply also to information retrieved from the web. The

	Research questions	Motive and relevance to the evaluation instrument
		quality of information retrieved by a search engine is indicative of the quality of the search engine itself.
2	Are there different kinds of search tools, and what are their distinguishing features? Are some more suitable than others, and are there general categories of search tools?	Different search tools will be studied to determine their characteristics and usability. Their characteristics and unique features will be used to identify possible categories and will inform the development of the evaluation tool. Ideally, an evaluation tool should suggest a benchmark for public/web based search engines.
3	What are the different parts of a search engine, and does this impact on its search efficiency?	Influences the choice of search tool, and thus the evaluation tool.
4	How does search engines relate to one another, and do they search the same part of the web?	Search engines searches often overlap, but not always. The size of a search engine might be important during some searches, and should therefore form part of an evaluation tool.
5	Are some search engines more popular than others and what are their distinguishing characteristics?	It is argued that the major search engines are considered major for specific reasons, and the characteristics of the major search engines would inform the evaluation tool development – using a 'best practice' approach.
6	How do search engines search for information, where do they find it, and how are the results presented?	Insight into how search engines search, their sources and how results are presented, as well as the rankings allocated to the results are critical factors

	Research questions	Motive and relevance to the evaluation instrument when considering a search engine's usability.
7	Is the number of records retrieved by a search engine important?	The recall/precision ratio should be considered when evaluating search engines.
8	What role does the experience and background of the user play when conducting a search?	Search proficiency and background knowledge of the user might influence the choice of search tool and its capabilities.
9	How does a user search for information using a search engine?	Differing user search methodologies and strategies affect the selection of appropriate search engines and their search results.
10	What are the drawbacks of using web based search engines – what are the difficulties of using them, and how can these difficulties be addressed?	A search engine evaluation tool will attempt to minimize drawbacks and optimize search capabilities.
11	Given the research into the questions set out above, is it now possible to design an instrument to evaluate search engines?	An effective search engine evaluation tool will be instrumental in selecting appropriate search tools for specific users and their queries.
12	How will local search engines perform when evaluated against the evaluation tool?	The evaluation tool will provide insight into the functionality and efficiency of local search engines, assisting potential users in search engine selection.

It is proposed that answers to these questions will indicate possible benchmarks for search engines and as such will be useful to design a search engine evaluation tool that can be used in evaluating search engines and to identify appropriate search tools for specific users and queries.

1.2 Goals

This study attempted to develop a methodology or an instrument to evaluate search engines to optimize search results. It is proposed that this evaluation methodology / instrument for search engines will enable researchers to identify the best search engine to obtain reliable, valid and current information faster.

1.3 Methodology

A literature study was conducted to research possible evaluation angles for Web search tools or search engines. To this end, major search engines were identified, described and compared not so much as to identify superior search engines, but rather to identify a range of desirable search engine characteristics and capabilities. The study also provides a background of information evaluation and information retrieval principles.

The literature was selected from a wide range of sources, both web-based and traditional. Because search engine development is a very dynamic field, it was attempted to, where possible, source the most recent information.

The information obtained by the search engine comparisons and descriptions was used to identify benchmarks of desirable search engine features and applied to develop a methodology / instrument for search engine evaluation. This instrument was applied to a limited number of local search engines to test the methodology.

2 INTRODUCTION

Information retrieval on the Internet and Web differs significantly from retrieval in traditional indexed databases and the way that people search for information has changed (Ding, Chowdhury & Foo, 2000). The search differences originate in the dynamism of the Web, its hyperlinked character, the absence of controlled indexing vocabulary, the heterogeneity of document types and authoring styles and the easy access that different types of users have to it (Gwizdka & Chignell, [1998]).

Conducting an online search for information *suspected* to be available, (it might not be) on the Internet can be time consuming and often frustrating. The dynamic, already monstrous Internet, its lack of control or official organization, and the fact that any one with the right hardware and software can publish information on the Internet, makes finding the proverbial needle in a haystack seem easy and attainable.

One can find information on the Internet in many ways. One could use WAIS, Archie, Veronica, Gopher, and FTP. All these preceded the WWW, but the Web has overshadowed them all. (Habib & Balliot, 2003). This study focused on finding information on the Internet via the World Wide Web.

The World Wide Web, also known as the Web or WWW, consists of a collection of documents stored on computers around the world. These specialized computers are linked to form part of a worldwide communication system called the Internet. When conducting an online Internet search, the user's browser (computer program that searches the Internet) goes to Web sites where documents are stored and retrieve the requested information for display on the computer screen. The Internet is the communication system by which the information travels (Habib & Balliot, 2003).

Although no one is in control of the information on the Internet, search tools greatly enhance the likelihood of finding the information required. Web search tools / engines first came into existence in 1994 (Chu & Rosenthal, 1996) and this study explores and studies these search tools and will attempt to develop an instrument or tool to help the user identify the best possible search strategy and search tool for a particular query.

The first chapters of the study provide basic background to the Internet, the Web, and the information available here, followed by a study and analysis of the different kinds of search tools, search methodologies, search engine selection, trouble shooting and more.

3 FROM ARACHNIDS TO FORMAL ATTIRE

For a long time, the World Wide Web (WWW) was likened to a giant spider's Web where all documents/information were somehow connected to everything else. One could move from one edge of the Web just by clicking on the appropriate interconnected string of hyperlinks. According to this "Small world" theory, every Web page is thought to be separated from any other Web page by an average of around 19 clicks (Laudon & Traver, 2002:18).

However, Laudon & Traver (2002:18) refers to recent research that found that the Web is not structured like a spider's web at all, but rather like a bow tie. Researchers discovered that the bow tie web has a '"strongly connected component" (SCC) composed of around 56 million web pages. (See Figure 1 below.) On the right side of the bow tie is a set of 44 million OUT pages that one can get to from the center, but one cannot return to the center from them. OUT pages are generally designed to trap the visitor at the site. On the left side of the bow tie is a set of 44 million IN pages from which one can get to the center, but that one cannot travel to from the center. These are referred to as newly created ''newbie'' pages that have not yet been linked to many center pages. 43 million pages were classified as ''tendrils'' - pages that do not link to the center and that cannot be linked to from the center. The tendril pages (or Dangling links) may be linked to IN and OUT pages. Tendril pages would also occasionally link to one another without passing through the center –referred to as ''tubers''. The researchers found also that there were 16 million pages totally disconnected from everything, called Orphan pages (WebWorkshop, 2003).

From arachnids to formal attire – the web is no longer seen as a spider's web, but rather as a bow tie:

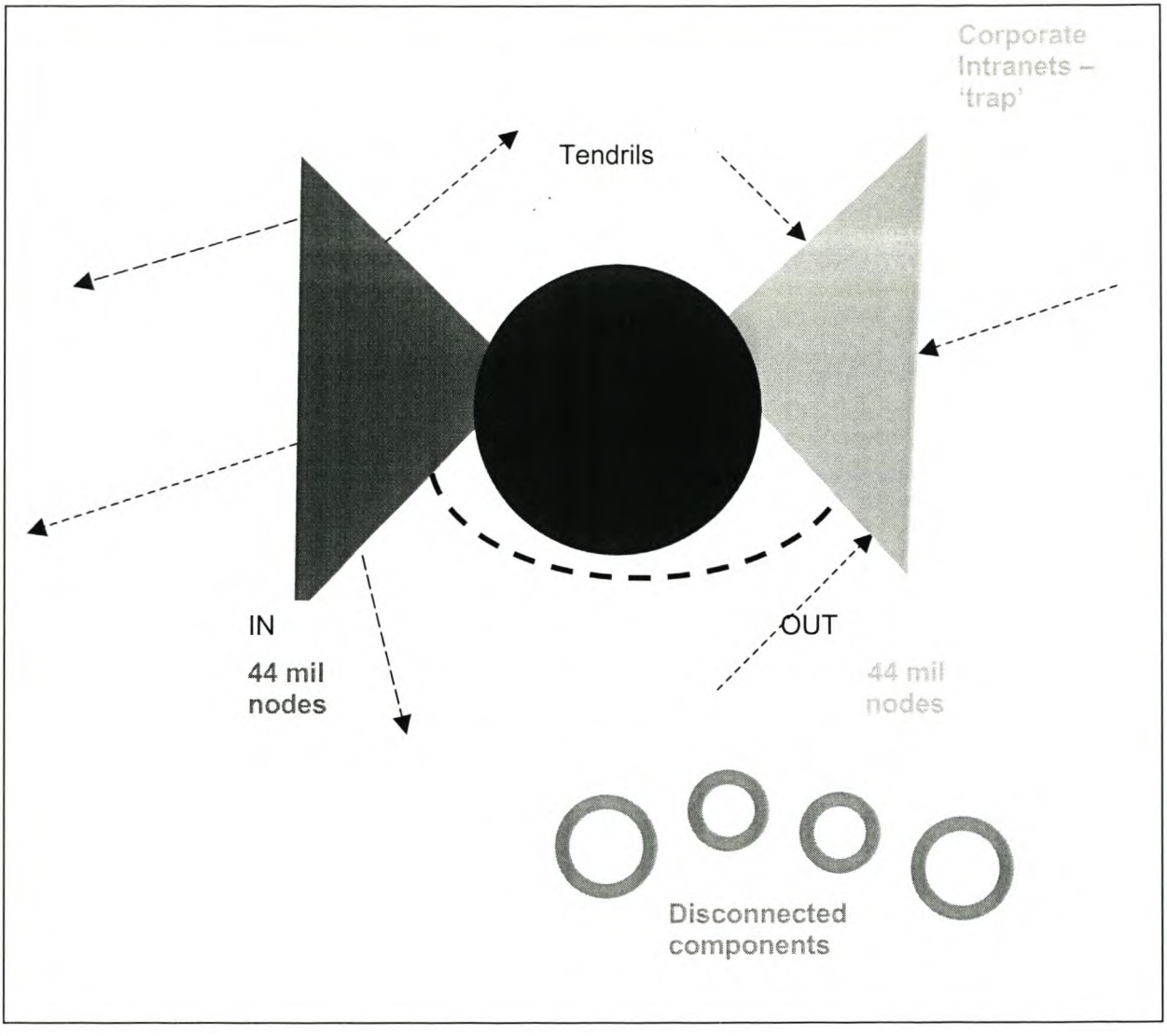


Figure 1: the Web as a bow tie

Research suggests that there is a 75% chance that there is no path from one randomly chosen page to another. This would explain why the most advanced Web search engines only index around six million web sites, when the overall population of web sites is over 70 million. The researches pointed out that most web sites could not be found by search engines because their pages are not well connected or linked to the central core of the Web.

Different users use Search engines for different reasons to find different kinds of information. Users must be able to find a Web site, and Web site managers should ensure that their Web pages are part of the connected central core of the Web. Web sites should have as many links as possible to and from other relevant sites – especially those sites within the SCC.

4 THE INVISIBLE WEB

Search engines and other search tools attempt to find and index as many sites as possible, but none can come close to indexing the entire Web, much less the entire Internet. Notess (2003b) list the following information items that are missing from search engine results:

- Content in sites that require a log in
- CGI output e.g. data requested by a form
- Intranets
- Pages not linked to from anywhere else
- Commercial resources with domain limitations
- Pages excluded by robots.txt file
- Content of some Adobe PDF and formatted files
- Non-Web resources: email lists, chat, IM, cookies, etc.
- Very current information, news, press releases
- Some multimedia files content: text in pictures, sound files, video files.

These "invisible" items of information are referred to as the Invisible Web.

Search tools that search (some) of the Invisible Web are discussed in Chapter 6: Different kinds of Search Tools. Further in-depth discussions and evaluation of search tools aimed at the Invisible Web falls outside the scope of this study.

5 EVALUATING WORLD WIDE WEB INFORMATION

Users employ web search tools to find relevant and adequate information for a specific query. The user expects or hopes that the search results returned by the search engine would be appropriate, yet this cannot be assumed. Internet publishing is not controlled, and the responsibility for reliable information retrieval ultimately rests with the user. Ideally, search engines should retrieve only appropriate and relevant information, so in studying search engine characteristics and features the quality of the information retrieved could be indicative of the quality of the search engine employed.

Because web sites are often a blend of information, entertainment and advertising that complicates evaluation, this chapter deals with evaluating retrieved information.

Ferguson (2003) and Indiana University (2003) suggest the following seven information evaluation criteria:

5.1 Authority

Anyone can publish anything on the Web. It is therefore important to identify the author of a Web document and verify the author's qualifications relating to the topic. The author should be identified and a link provided to access more information, such as his/her credentials. Ideally, a physical address and telephone number for the author and/or the institution should be provided on the home page. Searching for similar documents using the author's name or organization as search terms might provide more information on the author or organization. If no author is mentioned, the publisher, institution or organization responsible for sponsoring the document's host Web site must be examined.

5.2 Publisher or sponsoring agency

The credentials and motivations of the organization or people responsible for maintaining a Web site could indicate credibility. The responsible organization must be clearly identifiable. Consistent headers, footers or wallpaper could point to association with a larger Web site. The Web page should provide a link to information about the publisher or sponsoring agency. By examining the URL (Uniform Resource Locator or Web address) one can see whether the page is a part of an official site or part of someone's personal account (a tilde ~ in the URL indicates a personal Web page). Credibility is improved if the information has been published in a respected (subject relevant) journal.

5.3 The URL

The URL provides more information about the type of Web site and the origin of the information, as well as indicate the intended audience of the site, e.g. educational domains indicates an academic audience, commercial domains might be targeted at existing and potential customers or clients, and a government domain indicates an official government site with information aimed at government officials, the citizens of the country or information for foreign visitors.

Components of the URL include:

- **Host computer:** the host or name of the computer server where a Web site is located. The computer name follows the "www".
- **Domain name:** the final few letters that follow the host computer name, e.g.
  - .edu – created at an educational institution (research – student pages)
  - .net – varies greatly, often indicates that the site was created by a person or group using an Internet service or network provider (services to subscribing customers)
  - .gov – created by a government agency (official government information)
  - .org – varies, usually created by a non-profit organization (may promote a specific point of view)
  - .com – commercial enterprise (may try to sell products or services)
  - .mil – created by the U.S. military
  - .in.us – created by state-supported institution of Indiana – the .us domain requires a state code as a second level domain, e.g. www.monroe.lib.in.us
- **Directory path:** the directory, in which the file is located, usually followed by a forward slash (/).

5.4 Bias/point of view

Information is rarely neutral. The user should identify biases in a document and decide about the wisdom of using the information contained within the document. The author's or editor's point of view should be clearly stated. Be aware that the information contained may influence the user's opinion of it, especially in a commercial environment. The document may come from a server that is sponsored by an organization with a specific agenda (political, commercial, or philosophical). If the information refers to a controversial issue, the author should acknowledge the controversy. Obvious bias in a document does not necessarily disqualify the information because various sources of information are appropriate for use in different research situations. Critical thinking and evaluation on the part of the user is important.

5.5 Accuracy and reliability

The Web page should provide a way for the user to verify background information provided in the document. Any grammar or spelling errors in the document should immediately raise suspicion about the content. If the document quotes or refers to other sources, it should provide a bibliography or links to the source of the documents mentioned. A research document should include the gathered data and the research methods must be explained.

5.6 Currency

Once a Web site is placed on a server, it will remain there until it is either removed or the server is turned off. Information on the Web is not necessarily up to date, and a web search provides only a snapshot of the page as it was when it was spidered. Consider spidering frequency as well as content updates. Currency may be important to the research topic, and the document should clearly indicate the currency of the source. The page may be dated – users should check whether the page is regularly maintained and updated. Often the contents will contain clues to the currency, e.g. 2000 Census. The sources used in the bibliography or references may also provide dates and clues. References to current news events as well as current links (i.e. no expired or moved links) may indicate recent maintenance and currency.

5.7 Purpose of the Web site

The user should determine the purpose of the Web site. It could be to inform about current events and new information, or to explain, e.g. teaching and instructing. The Web site may also try to persuade, e.g. to change the user's mind or to sell something. Most Web sites fall into these broad categories:

- Advocacy Web pages
- Business / marketing Web pages
- Information Web pages
- News Web pages
- Personal Web pages.

6 DIFFERENT KINDS OF SEARCH TOOLS

Before one can start to examine search engine methodology, it is prudent to explore the different kinds of search tools, generally referred to in the literature as search engines.

Search engines provide a form for entering search phrases that return a list of results. The results generally include a mixture of data from directories, spiders and paid listings (Search engine marketing, 2003b). The search results are also ranked in some order based on a scoring criteria (Haynal, 1999) or page ranking system. Each hit or search result displayed includes a hyperlink to the Web page and a portion of the Web page's content. The URL might provide clues as to the 'hit fit' – an appropriate and fitting answer to the query. Results are presented in batches of 10 – 20 results. After working through the top ranking results, one can search through the next batch.

Search engines profit by selling advertising on their Web site. Search engines gather information into a data bank, which is a collection of non-relational data such as different Web pages (What is a search engine? 2002).

It is a challenge to define and categorize search engines. Different authors have different ideas of what actually constitute a category, and which are sub-categories. Some search engines fit comfortably into more than one category.

The following sections explore the different kinds of search tools, or generally called search engines, described in the literature studied.

6.1 General purpose, standard or automatic search engines

Search engines are useful to find unique keywords, phrases, quotes and information embedded in the full text of Web pages, and to source a wide range of responses to specific queries (Zaino, [2003l]). Search engines conduct keyword searches, as opposed to subject searches conducted in directories.

A search engine allows users to find Web pages or sites that contain a given keyword and/or phrase. When a user enters and makes a query, the engine returns the pages on the site that it determines to best match the search criteria. Each search engine has its own method of indexing a page or site to determine which result ranks highest (Level Ten, 2002c).

Search engines match terms, phrases or words with sites listed in its database. Most will default to an implied OR, which means that they produce many hits. Most rank the hits according to relevance. Because each search engine is unique, it is important to match the best search engine for a given topic (JMU Libraries, 2003).

Different to directories, search engines are unmanned. The process of indexing sites is automated and the human component is completely removed. A software program (known as a robot, spider or crawler) reads or indexes the Web pages, follows links between pages and sites, and collects information to store for later use. The spider will automatically return to the same site periodically to check for new content or new pages. The results of this spidering are saved in the index of the engine and serve as the basis to orient each search query. Given the automation process and the size of the Internet, these devices grow to upwards of 250 to 500 million pages. These efficiencies enable the search engine to cover a variety and number of sites. Each page stored in the database directory is ranked based on the contents of each Web page, including the title of the page, meta tags, text, images, etc. The index entries will change with any alterations and updates on a particular page and Web site following a new visit by the spider (The Web pros group, 2004). Page titles and body copy are essential elements in search engine listings (Carr, Santowski & Marzolf, 2000).

A search on a search engine means that the search engine's index is being searched. The search results will depend on the contents of the index, which in turn are based on the contents of each Web page that was spidered. A drawback is that the information quantities may be voluminous compared to a directory-based site. This search tool is useful to source very specific, current information. Robot sites also offer current links to other sites as referenced in the results (Day, 2001).

Examples: Google, Excite, and HotBot

6.2 True search engines

True search engines, just as standard/automated/general search engines crawl the Web, but then people search through what they have found to compile the index. There is no human interference in standard search engines.

Example: AltaVista

6.3 Global and local search engines

Global search engines are search engines with general, broad coverage of Internet information. Efficiency and efficacy are the major concerns (Search engine types, 2002). Some search engines are designed around a desired demographic, e.g. in the UK most versions of search engines filter results slightly differently from their .com parents to weigh Web sites with a UK domain (.co.uk) more favorably. They may also favor .com domains that are hosted in the UK, or are linked to by mainly UK sites. Directories sometimes maintain a separate database. Some databases, such as Inktomi, are common across the world, and used by many US and UK portals (Spannerworks, 2003a).

The three South African search engines studied in Chapter 17 also tended to weigh local web sites more favorably.

6.4 Dedicated search engines

(See also Portals, directories and specialized directories or subject guides.)

Dedicated search engines focus on a particular sector or industry such as people or business (Search engine types, 2002).

6.5 Natural language search engines

Natural language search engines allow for entering and searching using natural language – plain English, as one would ask a question, without the application of any special syntax such as Boolean searching (TRC, 2002d).

Examples: AskJeeves, Ask Jeeves Kids, and AltaVista Canada.

6.6 Filtered search engines

Filtered search engines allow for child safe searches. The entire Web can be searched or search only a selection of pre-approved sites will be searched. Results are filtered to remove possible objectionable material (TRC, 2002a).

Examples: Cyber Guide, Cybersleuth Kids, Awesome Library, and AltaVista – turn on family filter.

6.7 Media search engines

Media search engines expedite searching for multimedia files such as sound, picture or movie files (TRC, 2002c).

Examples: AltaVista – Image Search, Audiofind, Ditto.com, The Amazing Picture Machine, WebSEEK, Pics4Learning, SingingFish, and http://search.mediasite.net, that searches for multimedia information and files.

6.8 Shopping search engines

Shopping search engines are dedicated to finding particular products that the searcher intends to purchase, or to compare prices and availability.

Examples of shopping search engines (Wall, [2003c]): BizRate, Buyer's Index, DealTime, Froogle, mySimon, NextTag, PriceGrabber, PriceScan.

6.9 Library Gateways

Library Gateways are also known as the ''Invisible Web". The Web contains thousands of databases maintained by universities, libraries, government organizations, and businesses that are inaccessible using standard search engines. Gateways provide access to database information. Library Gateways contain subject directories of reviewed sites. Subject specialists create these directories to support research needs and to highlight high quality sites on the Web. Gateways offer access to specialized databases (Zaino, [2003c]).

Library Gateways can be used to search for high quality information sites and when a subject specialty database is required. Library Gateways are useful when searching for research and reference information (Zaino, [2003k]).

Examples: New Canaan Library (www.newcanaanlibrary.org), Internet Public Library (www.ipl.org), Digital Librarian (www.servtech.com/~mvail), Living Web Library (www.livingWeb.com/library/search.htm)

6.10 Directories, Subject Trees and Subject Guides

Hierarchical search engines or subject directories have links that have been screened, selected and catalogued in the engine's own database. When searching one of these sites, only one search engine's particular database of links, considered worthy of inclusion, is searched. These search engines are considered safe for children and simple to use (TRC, 2002b).

Directories classify the Internet according to a knowledge scheme, useful for a broad overview of a specific subject area. Usually the classification scheme is unique and lower level subjects may appear in unusual headings. They may also contain a search function that allows simple location

(JMU Libraries, 2003). Directories are collections of human reviewed Web sites that have been described and categorized by subject or location, e.g. dir.yahoo, dmoz.org (Search engine marketing, 2003b). Directories provide a service similar to the traditional Yellow Pages. The Open Directory Project is considered to be one of the most important directories as a number of key search engines use it as one source of input to their spiders, and the category gives them a basis on which to initially index new Web sites (Crickett Software Limited, 2003). Directories conduct subject searches, as opposed to search engines that conduct keyword searches.

From an initial menu, the searcher must decide the general location of the subject searched. As more specific menu items are selected, the searcher is taken deeper into the menu hierarchy until eventually the bottom page in the hierarchy is reached. The bottom page contains the hyperlinks that reroute the searcher from the directory to the specific Web sites containing the desired subject (Haynal, 1999). These databases are comparatively small and the frequency of the updating is relatively low (Day, 2001).

A directory may offer a search option, but this is not a search of the Internet, but a keyword search of the Web pages contained within the directory. Keyword searches on this search option should best be limited to a subject search (Haynal, 1999). Directories are useful resources to learn the terminology of a subject field.

People create the Web site's listing on the search page create directories, as opposed to robots or spiders that does this automatically for search engines. A short description and the URL of the Web site are submitted to the directory, or editors write one for sites they review. Upon approval, the search directory assigns the Web site to an appropriate category within the large Web site. The Web pros group (2004) argues that directories provide more targeted results than search engines. A search on the directory site looks for matches only within the descriptions submitted, and not information found on the Web pages. Most directories allow only one submission per Web site (IWD, 2002b).

Directories have the advantage that humans are generally better equipped to identify interesting, worthwhile or relevant sites than software are, and directories are considered authorities by other search engines (Spannerworks, 2003b). On the other hand, humans are slower than software crawlers are; they are not as accurate and may use very subjective criteria to categorize sites.

Clearly, the accuracy, scope and depth of the site description submitted to a directory are vital to improve the hit rate of a particular site.

Examples: Yahoo (dir.yahoo), Open Directory Project (www.opendirectory.com), AOL, AltaVista, www.nbci.com, Looksmart, Infoseek, DMOZ, Yahooligans, PINAKES and About.

6.11 Meta search engines or multi-engine searching

A search form allows users to get results from a number of different search engines at once, e.g. Dogpile.com (Search engine marketing, 2003b). Meta search engines search the various search engines from a single site using the same interface. They offer a quick way of determining which search engines retrieve the most relevant results for a search (JMU Libraries, 2003). Meta search engines are useful to do comprehensive searches on a very specific topic.

A common or natural language request is electronically and transparently distributed to multiple search engines. The relevant hits are integrated into a single, ranked list. The protocol is tolerant of imprecise search questions or keywords (Day, 2001).

Many meta search engines use a combination of methods to generate their results. They might include or combine automatic search information and directory information, or they may use the directory information to sort or filter the automatic search results, or they may combine pay per click results with their own results (The Web pros group, 2004). www.infind.com removes redundancies and clusters the results into understandable groupings (IWD, 2002b).

A cluster is a grouping of representations of similar documents. In a vector space model, one can perform retrieval by comparing a query vector with the centroids of a cluster. One can continue search in the most promising clusters. A vector space model is a representation of documents and queries where they are converted into vectors. The features of these vectors are usually words in the document or query, after stemming and removing stop words. The vectors are weighted to give emphasis to terms that exemplify meaning and are useful in retrieval. During retrieval, the query vector is compared to each document vector, and those that are closest to the query are considered similar, and are returned. Weighting, when referring to terms, is the process of giving emphasis to the parameters for terms that are more important. In a vector space model, this is applied to the features of each vector. Boolean is a popular weighting scheme, or term frequency alone could be used (Weiss, 1997).

Searching using meta search engines increases the error margin substantially (Search engine types, 2002) and they are prone to time-outs when search processing takes too long (Notess 2002a). Habib and Balliot (2003) disagree, and describe the hit list of meta search engines as

more likely to be short and of a high relevance. Generally, the best use of meta search engines is to determine whether something can be found on the Internet (KCPL, 2002c).

Examples: MetaCrawler, Ixquick, Dogpile (uses Yahoo, Thunderstone, Lycos' A2Z, GoTo.com, Mining Co., Excite Guide, PlanetSearch, What U Seek, Magellan, Lycos, WebCrawler, InfoSeek, Excite and AltaVista), www.mamma.com (uses Yahoo, Lycos, InfoSeek, GoTo.com, FindWhat, MSN, AskJeeves and NBCi), www.infind.com (queries WebCrawler, Yahoo, Lycos, AltaVista, InfoSeek and Excite), SavvySearch, The Big Hub, and Vivisimo.

6.12 Portals

Portals are sites that specialize in reaching a particular audience. They attract a regular repeat audience made up of mostly users with the same interest that the portal caters for (Crickett Software Limited, 2003). Portals are entry points where users can check the weather, conduct a Web search, browse directories, view stock quotes, etc. (Ayache, 2003). Google, Yahoo and AltaVista are examples of portals that contain search engines. Portals can often be customized according to the requirements of the user.

Examples: MSN.com, Yahoo.com.

6.13 Pay Per Click (PPC) / Ranking / Placement or Position search engine

PPC (2002) calls this the evolution of the 'traditional' search engine. Paid placements are recent additions to search engines. Paid placement services were developed to weed out the true businesses from the plethora of nonsense Web sites that had popped up in top listings. Paid placement usually involves selecting a target key word or phrase that applies to a business and bidding for a price per click that they will pay to be ranked at the top of the page. Business buy advertising for particular key words - those used by Google is a good example (Crickett Software Limited, 2003). Depending on competition, this is useful to get good placement (Ayache, 2003). When a Web searcher searches for that particular key word, that advertiser's listing will be ranked according to the amount bided, compared with all other bids on that key word. Regardless of how many times a listing appears, that advertiser only pay when a searcher actually clicks on their listing and visits their Web site.

With traditional search engines, advertisers submit their Web sites, which are then indexed and ranked, based on different factors. This is time consuming and may take months before results are seen. There is also no guarantee of top placement within the search results.

With PPC search engines, the process is faster and advertisers can determine the placement of their listing within their search results, depending on the price they are willing to pay.

Web users and researchers should take note of PPC search engines and the implications for ranking of results – the advertiser that paid the highest price for a high ranking might not necessarily provide the best information. A high price ranking does not equate a high relevancy ranking.

There are guidelines and rules to prevent advertisers from manipulating the results and minimum bid amounts and deposits are required (The pros group, 2004).

Search results from many other search engines also include pay per click results. The user using a Pay per click search engine must consider its functionality as a search engine as well as its user friendliness, speed, possible distracting interfaces and accuracy of navigation bars.

Examples: Looksmart, FindWhat.com, Overture, Google AdWords, Sprinks, Ah-Ha.com, Kanoodle, and Godado.

6.14 Free listings

Most common search engines offer free listing to new Web pages. The drawback for these new sites is that there is no guarantee that the search engine will spider or index the new site (Crickett Software Limited, 2003).

6.15 Clustering search engines

Clustering search engines automatically group search results into related themes, thus focusing results for ambiguous terms. A search for 'blues', for example, might present the themes 'music, 'Oxford rowing team', and 'depression'. When selecting a theme, only results in this area will be displayed (Spannerworks, 2003b).

Examples: Vivisimo and AltaVista (using the clustering search engine Prisma).

6.16 Subject guides

(See also Hierarchical search engines and Directories)

Subject guides are used when browsing, for more generalized searches, and work well for popular topics, organizations, commercial sites and products (Zaino, [2003m]). Subject guides are created and maintained by human editors. The editor reviews and selects sites and compiles directories based on previously determined selection criteria. Their listed resources are usually

annotated. Subject guides tend to be smaller than search engine databases, typically indexing the home page or top-level pages of a site. They may include a search engine for searching their own directory or the Web, should a directory search provide no results (JMU Libraries, 2003).

Subject guides return fewer out of context results, they deliver a higher quality of content and they are organized into browsable subject categories. Unfortunately, they also present more dead links and have smaller databases.

Examples: About.com, Ask Jeeves, Magellan, and Snap.

6.17 Virtual libraries

Virtual libraries are subject-specific indices that are created and maintained by people that are interested in that particular topic. The virtual library resembles Yahoo, but the individual subject areas are distributed and maintained by subject experts. The www.VirtualLibrary provides links to a list of virtual libraries (Haynal, 1999).

6.18 Intelligent agents

Intelligent agents are software robots that carry out a task unsupervised and apply some degree of intelligence to the task, e.g. an agent that searches the Internet for interesting material can be told by the user whether what it found was interesting or not. In this way, it can be 'trained' to be more successful in the future. Some intelligent agents can also interact with one another (CompInfo, 2002).

Intelligent agents transform passive search and retrieval engines into active, personal assistants. The combination of effective information retrieval techniques and autonomous intelligent agents can improve the performance of short-term retrieval in an existing search or retrieval engine (Jansen, 1996).

Agents for use on the Web allow users to conduct tasks on the Internet faster and more efficiently. Intelligent Agents ([date unknown]) lists the following types of Bots/Intelligent Agents: Shopping Bots, Search Bots, Tracking Bots, Download Bots, Surf Bots, Games Bots, Web Development Bots, and Artificial Life Bots.

Example: WebSeeker (http://www.bluesquirrel.com/products/seeker/) – it combines the results of multiple search engines, delivers a clean list of results that can be saved, viewed offline, easily organized and updated automatically.

6.19 Specialized directories/databases

The following specialized databases are useful when one knows exactly what one is looking for:

- Newsgroups, online discussions, e.g. Deja News Service, AltaVista (select 'Usenet' instead of 'Web'), Onelist.com
- Internet marketing, starting your own business, e.g. Internet Marketing Center, Ad Resources
- People, Roadmap, Internet Address Finder, Switchboard, Yahoo People Search, AltaVista People Finder
- Company information, e.g. Inquiry.com, Hoover's Online, PR Newswire, LEXIS-NEXIS
- News, e.g. Pathfinder's News Now, CNN Webspace Search Engine, The New York Times on the Web
- Jobs, e.g. Yahoo' s Employment Ads
- Friends, lovers and matchmaking, e.g. Yahoo Personals, Match.com.

6.20 Hybrid search engines

Powerful search engines that combine the best features of spiders and directories to obtain information and organize it into conceptually related fields (IWD, 2002b). The directory with search engine uses both the subject and keyword search. The tool initially follows a directory path for the keyword search. It then progressively narrows the search field with repeated visits to a selected site. This a good approach if the searcher is not familiar with the correct subject or keywords to use (Day, 2001).

Examples: Excite, Infoseek, www.go.com.

6.21 Other search engines

IWD (2002b) classify the following search tool as "other":

- www.myivan.com Speech driven engine that requires downloading a large file.

7 SUGGESTED CATEGORIZATION OF SEARCH ENGINES

From the previous chapter it is clear that there are about as many types of search engines or search tools, as there are people willing to categorize them. This study will distinguish between search engines and search tools. This study will regard 'search tools' to be the overhead, general term to include search engines, directories, shopping search engines, hybrids, etc. as referred to in previous chapters.

As search tools benefit from developments in software and hardware, they are increasingly hard to categorize, and the boundaries between different types are increasingly blurred.

From the long list of search tools they were initially grouped together in the following six categories:

- Pure search engines
- Pure directories
- Hybrid systems
- Virtual libraries
- Specialized directories, and
- Intelligent agents.

This list, however, still seems to be too cumbersome and general. For the purpose of this study, it was decided on the following categorization for search tools:

- Search engines
- Directories
- Portals / hubs ("evolved" hybrid systems) and
- Intelligent agents.

Some search tools may fall into more than one of the new categories. See Table 1.

Table 1: Search tool categories

SEARCH TOOL CATEGORY				
Search Engine (SE)	**Directory**	**Portal / Hub Specialized**	**General**	**Intelligent Agent**
FEATURES				
Very specific information query	Browsing or for very in-depth information on a well known topic	Focusing on a specific topic, field of study or business sector	Acting as a jumping off point for users	Web use: users can conduct tasks faster and more efficient
Browser window to type in and refine search	Hierarchical search	Personalized features: selected information or report displays & notifications	Often a hybrid of a search engine and directory, with similar features	Transform passive search and retrieval engines into active, personal assistants
Ranked results	Human edited		Personalized features: email and specific, selected, information displays	Can improve the performance of short-term retrieval in an existing search or retrieval engine
Uses spiders, Web pages are indexed			Displays general interest information, e.g. news, weather, shopping opportunities, etc	
EXAMPLES:				
AltaVista	ODP / DMOZ	Deja News Service	Yahoo	Blue Squirrel

"New Classification" Search Engines	**"New Classification" Directories**
INCORPORATING:	
General purpose SE	Hierarchical SE
Standard SE	Library gateways

"New Classification" Search Engines	**"New Classification" Directories**
Automatic SE	Subject trees
True SE	Virtual libraries
Global & local SE	Subject guides
Natural language SE	Specialized directories
Filtered SE	
Media SE	
Shopping SE	
Meta SE	
Multi engine searching SE	
Pay per click SE	
Pay per ranking SE	
Pay per placement SE	
Pay per position SE	
Paid for listings SE	
Free listings SE	
Clustering SE	
Dedicated SE	

8 PARTS OF A SEARCH ENGINE

Carr, Santowski and Marzolf (2000) identify three major elements of search engines – spiders, indexes and search engine software.

8.1 Spiders, crawlers, robots or bots

Spiders are the workhorses of the Internet (Ayache, 2003). A spider is a program that crawls from page to page and indexes the content into large databases that can later be queried by search engines, e.g. GoogleBot, Inktomi, FAST (Search engine marketing, 2003b). Level Ten (2002d) defines a spider as a program that browses (crawls) Web sites extracting information for search engine databases. Spiders read the meta tags, pieces of information (invisible to the searcher), that are coded in the HTML of a page that describe the contents of the page (Spannerworks, 2003b). Spiders can be summoned to a site through search engine registration. They will also eventually find a site by following links from other sites – if there are any links from other sites. Spiders do not read sites as browsers do. Generally, they are unable to execute JavaScript, including links performed by scripting or frames links, or index content in images, and are thought of as very primitive browsers (Spannerworks, 2003b). Spiders explore sites by using hyperlinks, but they will only go so many levels deep and a visiting spider may not index an entire site. Web site designers can also block spiders from certain pages on their site that may contain sensitive or confidential information. When a spider discovers a new site, it sends information back to the main site to be indexed. Because Web documents are one of the least static forms of publishing, robots also update previously catalogued sites. The update frequency and comprehensiveness varies from one search engine to another (Barlow, 2002b).

Examples of search engine spiders: AltaVista, HotBot, Lycos, and WebCrawler.

8.2 Index

Contains a copy of every Web page that the spider finds. The index is updated with new information when a web page changes. It may take a while for new pages to be added to the index. A Web page may have been spidered, but not yet indexed. Until it is indexed, i.e. added to the index, it is unavailable to searchers using the search engine (Carr, Santowski & Marzolf, 2000).

8.3 Search engine software

This is the program (apart from the spider and indexing program) that sifts through the millions of pages recorded in the index to find matches to a search and rank them in order of what it believes to be most relevant (Carr, Santowski & Marzolf, 2000).

9 SEARCH ENGINE RELATIONSHIPS

9.1 Sharing results

Most of the major search engines relate to the others through the various sources they use and share (Ayache, 2003). This relationship between search engines is very dynamic. Different search engines purchase or use different directories and vice versa. Search engines are becoming more 'hybrid', and behave more like portals than ever before. This chapter examines this relationship between search engines. It is by no means regarded as fully up to date – it only reflects the situation as at a given time.

The following table was compiled by Le Roux (2003a) in the Search Engine Yearbook v. SEY 2003, and reflects search engine relationships for January 2003:

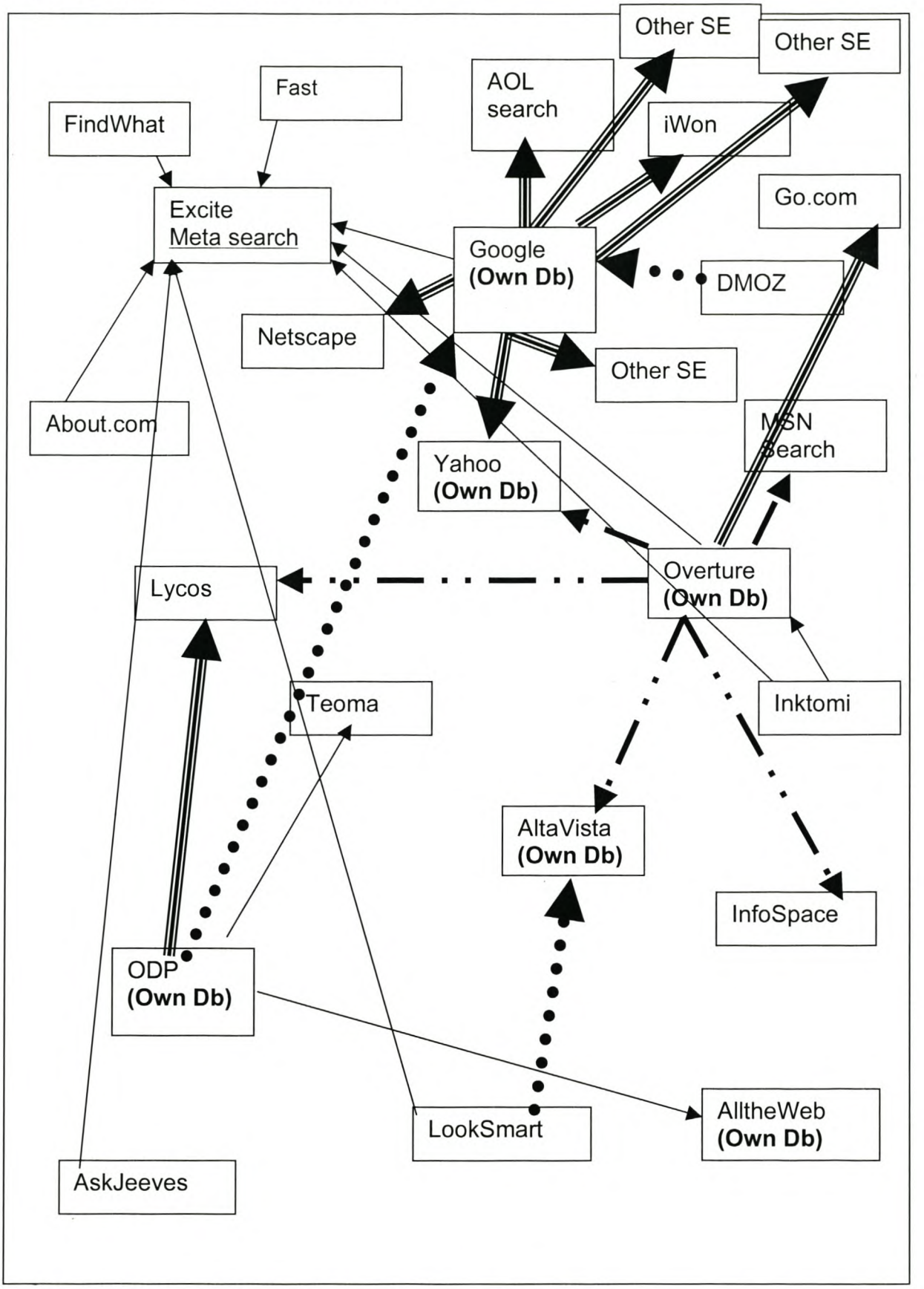
Table 2: Search engine relationships

Search engine	Receives results from	Sends results to
Google	Own database Directory listings from DMOZ	Main results to Yahoo, Netscape, iWon and AOL Search (and many smaller search engines). Paid listings (from AdWords) to Teoma, Netscape, AskJeeves and AOL Search
Yahoo	Own database Main results from Google Paid listings from Overture	None
AltaVista	Own database Directory listings from Looksmart Paid listings from Overture	None
ODP	Own database	Main results to Lycos Directory listings to Google Some results to AlltheWeb and Teoma
Overture	Own database	Main results to Go.com

Search engine	Receives results from	Sends results to
	Some results from Inktomi	Paid listings to Yahoo, MSN Search, Lycos, AltaVista and InfoSpace
Excite	Meta search. Receives results from Google, Looksmart, Inktomi, AskJeeves, About, Overture, FindWhat, Fast	None
AlltheWeb	Own database	None

Figure 2 represents the search engine relationships as in the table above, and clearly shows the interrelated nature of search tools:

Figure 2: Interrelated nature of search tools

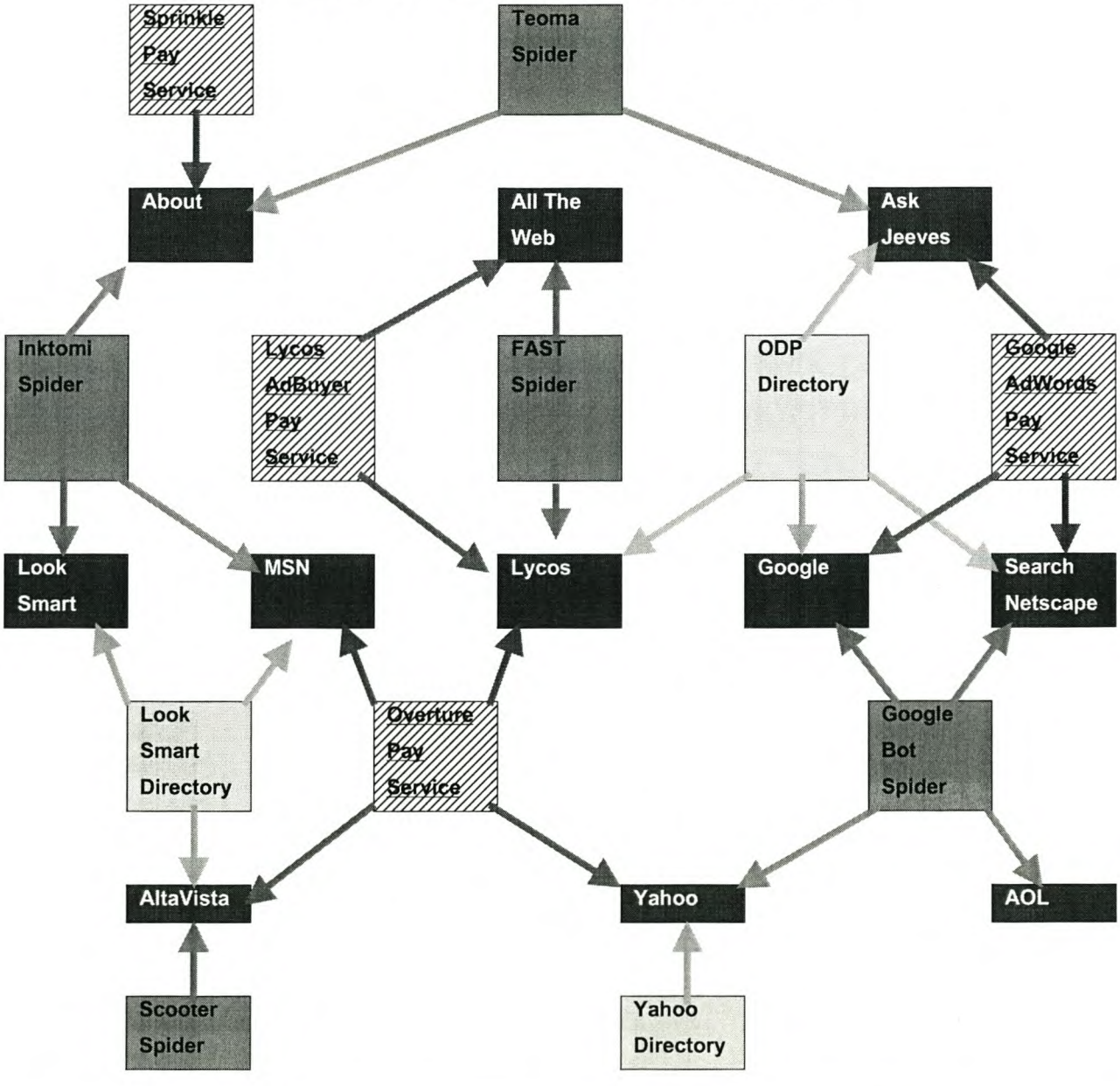


9.2 Search engine relationship: sharing of sources

Figure 3 depicts the competition between search engines for sources. It shows that sources reach not more than four portals (Search engine marketing, 2003a).

Legend:
11 Portals,
3 Directories,
5 Spiders, and
4 Paid services

Figure 3: Search engine competition for sources



9.3 Search engine speed

The following figure by Le Roux (2003b) shows the search engines' relationship with regard to speed – the response time of each search engine was divided by that of the fastest search engine (Google). The numbers are not response time in seconds, but response times relative to that of Google. All the response times of the search engines were actually very fast.

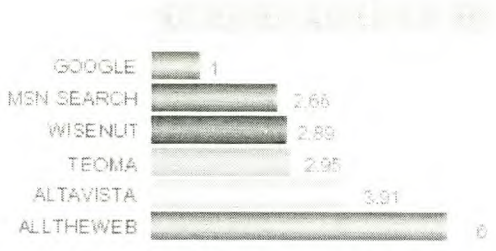


Figure 4: Search engine speed

From the Search Engine Yearbook v. SEY 2003 (Le Roux, 2003b)

Le Roux is of the opinion that these figures do not mean much and that any search engine that is significantly slower than the average search engine would just not be used by users, and would therefore be eliminated almost by default.

9.4 Search engine freshness

Although all search results are pictures of the past, it is important to know which search engine has taken the picture most recently. Results from a study done by Notess (2003m) show that most search engines have some results indexed in the last few days, but that the bulk of most of the databases is about one month old. Some pages may not have been re-indexed for much longer. Notess' study shows that MSN and HotBot had the best freshness average of 4 weeks old. Table 3 shows their results as found in the study:

Table 3: Search engine freshness

SEARCH ENGINE	NEWEST PAGE FOUND	ROUGH AVERAGE	OLDEST PAGE FOUND
MSN	1 day	4 weeks	51 days
HotBot	1 day	4 weeks	51 days
Google	2 days	1 month	165 days
AlltheWeb	1 day	1 month	599 days
AltaVista	0 days	3 months	108 days
GigaBlast	45 days	7 months	381 days

Table 3: Search engine freshness

SEARCH ENGINE	NEWEST PAGE FOUND	ROUGH AVERAGE	OLDEST PAGE FOUND
Teoma	41 days	2.5 months	81 days
WiseNut	133 days	6 months	183 days

Ideally, indexes should be updated daily, and that date must be reported.

9.5 Database size

Search engine size matters because a search tool will not find a document or record that does not exist in its database. In a dynamic Web environment, the large search engine database is a critical tool to find information other than the very general and popular content offered by portals. Larger databases are particularly important in the following cases (Notess, 2003o):

- Searching for plagiarism
- Name searches, especially uncommon names
- Citation verification
- Unusual, very specific, topics and
- Hard to find products.

Search engines all use different counting mechanisms, and the data base sizes should be seen as relative and approximate only.

Le Roux (2003b) conducted a study in the fourth quarter of 2002 to determine the different data base sizes of search engines. Le Roux warns that the results are from their own studies, and was not confirmed by the search engines. He regards these data base sizes as unofficial. The estimated values are the average of the reported database size at the time, the estimated database size reported on SearchEngineShowdown.com as well as their own estimates. Search engines typically spread their databases over several servers and many might have been unreachable or down for maintenance at the time the study was conducted.

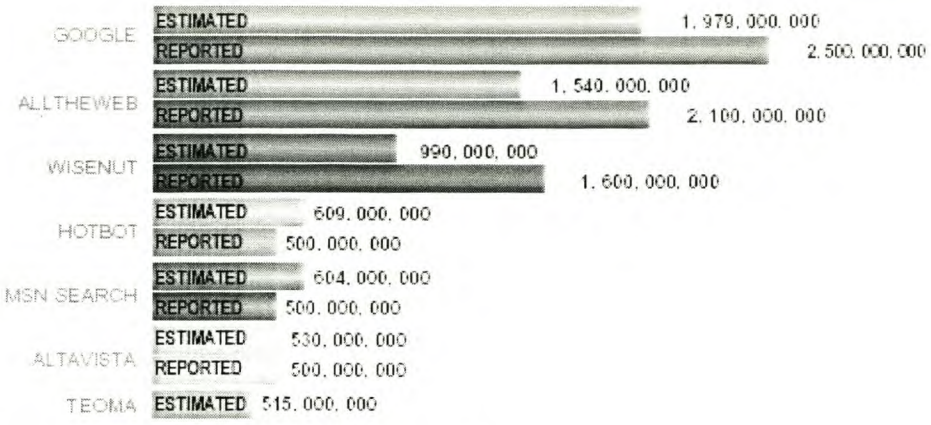


Figure 5: Search engine database size
From the Search Engine Yearbook v. SEY 2003 (Le Roux, 2003b)

9.6 Relative database size as compared to Google

Le Roux's (2003b) study was conducted in the fourth quarter of 2002. The values reflected are not indicative of actual database sizes, but show the sizes of some of the major search engines relative to the Google database. Le Roux describes the methodology in more detail in the Search Engine Yearbook.

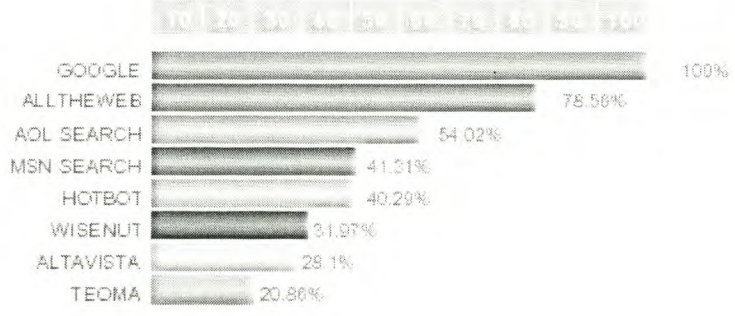


Figure 6: Relative database sizes compared to Google
From the Search Engine Yearbook v. SEY 2003 (Le Roux, 2003b)

These search engine relationships and source sharing emphasize the importance of selecting appropriate search engines for a specific query – search engines that do not overlap and just search the same sources and databases, but a combination of search tools that will cover different sources – casting the net of information wider than just the obvious.

10 MAJOR SEARCH ENGINES

Some search engines are regarded as major because they are either well known and/or well used. It also implies that these search engines generally generate more dependable results, are more likely to be well maintained, updated and upgraded to keep up with the growing Web. Identifying the major search engines is a relative process that is also subject to search engine changes and fashions.

The following lists of major search engines were obtained from the studied literature. The lists are by no means complete, and each differs slightly from the rest. From these different lists, (and some not detailed here) it will be attempted to compile a list of top five major search engines to use in this study.

Notess' (2003n) top five search engines for professional searchers are Google, AlltheWeb, AltaVista, Teoma and Inktomi (MSN Search and HotBot).

He lists the following search engines as the top ones for non-professional searchers: Yahoo, MSN Search, Google, AOL, AskJeeves, Lycos, and AltaVista.

IWD (2002a) lists the following search engines as major search engines: Yahoo, AltaVista, Excite, Google, iWon, Fast, Lycos, InfoSeek, WebCrawler, HotBot, MSN Search, Goto.com.

Search engine marketing's ([2002]) list: Google, AOL.com, Yahoo, MSN, and AskJeeves.

Spannerworks (2003a) argues that some search engines are more important than others in terms of delivering traffic. Google is regarded to be the most important search engine to be listed in, but so is MSN (that displays LookSmart) in the UK. Yahoo and Inktomi (contributing to, amongst others, HotBot, MSN and BBCi) are highly commended. The Open Directory Project is the biggest directory while Yahoo is the best known (Spannerworks, 2003b).

Barlow (2002b) lists the following well-known search engines: Yahoo, Google, FAST, Excite, AltaVista, Lycos, and HotBot.

Wall ([2003b]) lists the following: Google, AlltheWeb, AltaVista, AOL, AskJeeves, Earthlink, Excite, GigaBlast, Go, HotBot, InfoSpace, Inktomi, Lycos, MSN, Netscape, Overture, Teoma, and Wise Nut.

Le Roux's (2003a) list of major search engines: Google, AltaVista, Yahoo, DMOZ (ODP), Excite, AlltheWeb, Teoma, Direct Hit, Wise Nut.

Identifying the major or largest or most popular search tools is unfortunately not a clear cut process, and much of the information is subjective or guessed at, and the statistics change over time, with new developments, new partners, new players and new trends.

This study attempted to identify the top most popular search engines at this time with the aim of closely examining their features, why they are considered major and / or popular, and generally, how they are optimized to provide search results. The author argued that this should enable her to identify desirable and undesirable characteristics of search engines that can be used in the evaluation instrument. This was by no account an exhaustive literature study, although it was attempted to consult as widely as possible, and to use the latest statistics available, specifically those in the following sources: Barlow, 2002b; Collier & Arnold, 2003; comScore, 2003; Gray, [date unknown]; Hawking, Craswell & Griffiths, [2000]; IWD, 2002a; Le Roux, 2003a; Notess, 2003n; Peterson, 1997; Search engine marketing, [2002]; Spannerworks, 2003a; Testing labs, 2000; Wall, [2003b]; and WSPOS, [2000].

Following the above, the following search tools will be deemed as the top ranking tools for the purpose of this study:

- Google
- AltaVista
- Yahoo
- Lycos, and
- Excite.

10.1 Features of the major search engines

The following tables provide descriptions of each of these major search engines in terms of search engine features, and are by no means considered fully up to date or complete.

10.1.1 Google and AltaVista

Table 4: Google and AltaVista features

FEATURES:	SEARCH TOOL / ENGINE:	
	Google	**AltaVista**
1. Databases	Own database of indexed Web pages along with another collection of URLs that have not been indexed (e.g. redirected URLs, pages protected by a robots.rxt file, page with access restrictions, etc.), image database, Usenet news database (Google Groups), News search, Catalog search, Froogle (shopping search), Page Rank version of the Open Directory, specialized subsets (government, university, Linux Apple/Macintosh & a Microsoft search)	Web database: own indexed Web pages, incl. PDF, directory: Open Directory (Formerly LookSmart), news: own crawled pages, Ads: from Overture, images: own crawled, audio & video: own crawled, AltaVista Shortcuts.
2. Strengths	Full text – searches entire HTML file, "cached" page, relevance based on sites' linkages & authority (PageRank), additional databases, Web page translation, number search. Displays no pop-up advertising. Choice of 88 interface languages, including Afrikaans, Xhosa, Zulu & Swahili. Translation service enabled – English translations for pages in Italian, French, Spanish, German, and Portuguese. Very sophisticated database yet provides an uncluttered and easy to use format. It has become the default search engine for many web sites.	Full text, translate option, and powerful search features, some unique, international coverage, interfaces and foreign language handling. A leading search engine that has one of the largest databases and most effective search systems. Serves as the default search engine for LookSmart & Britannica Internet Guide.
3. Weaknesses	Limited search features: no nesting or truncation, does not support full Boolean search, link searches must be exact & are incomplete, only indexes 1st 101 KB of a Web page & about 120 KB of PDFs, site clustering is difficult to	Inconsistent, database not as large as it used to be, indexes only first 110K of a Web page and 750K of PDFs, no cached copies of pages or other file types beyond PDFs. Does not indicate the total number of related documents found. Running the exact same

FEATURES:	SEARCH TOOL / ENGINE:	
	Google	**AltaVista**
	turn off.	search will result in varying numbers of hits. Will produce masses of irrelevant hits if used properly.
4. Default operation	AND, phrase matches rank higher	AND Phrase search, removes punctuation marks or symbols, replaced by a space
5. Boolean searching	OR, - automatic AND, unable to nest operators, does not support the NOT operator	AND, OR, nesting, AND NOT, +, -, use some symbols: & for AND, | for OR, ! for AND NOT, ~ for NEAR
6. Stemming	Enabled, but searches are unable to select or de-select stemming – it is automatic, and may lead to frustration when precision searching.	
7. Proximity searching	Detects phrase matches even when the double quotes are not used, phrase matches are ranked higher, no other direct proximity searching available	Phrase, NEAR (within 10 words), within, before, before near, after, after near. Undocumented commands: numbered proximity, no order: within #, ~~ #, order (before): <, order, (before) & proximity: <~, order (after): >, order (after) & proximity: >~.
8. Truncation	Only in phrase searching. Represent single word within phrase, "trick" in using a wildcard word: use the asterisk * within a phrase search to match any word in that position, the use of multiple asterisks in this kind of search is supported. A tilde ~ directly before a search term searches for synonyms of the term.	Enabled, represent single word within phrase, use *, done after three letters for unlimited extra characters, internal or end truncation, also in phrases, not effective with numbers, effective with diacritics
9. Case sensitivity	No	Yes (advanced), no (simple), phrase retrieve exact matches
10. Field searching	intitle:, allintitle:, inurl:, allinurl:, link:, site:, related:, info:, define:, stocks: ticker symbols, allinanchor:, related: [URL] invokes GoogleScout to find pages similar in linkage patterns to the given URL	title:, url:, link:, host:, anchor:, image:, applet:, domain:, text: (terms somewhere other than image tag, link or URL), like: (use with complete URL), filetype:
11. Limits	Date (specific options provided), 34 languages, domain, filetype, adult content, and occurrences.	Date (user specified), 25 languages, region, file type, related pages, more pages from this site.
12. Stop words	No stop words excluded if + used or in phrase, stop words excluded if + not used or not in phrase, if a + is placed in	No stop words excluded (advanced), use quotation marks around stop words in simple search to force searching

FEATURES:	SEARCH TOOL / ENGINE:	
	Google	**AltaVista**
	front of a non-stop word in the same query, all + signs will be ignored. Exclusion of common words as per search is indicated on the results page below the search box.	
13. Sorting	By relevance as determined by its PageRank analysis, greater weight given to authoritative sites, clustering by site	By relevance, clustering by site
14. Display	Home page is user friendly and not overly busy. Uses link analysis to rank the pages displayed. Shows title, URL, two lines of text with search terms highlighted, file size in bites, file format, link to "cached page", "similar pages" link, most relevant results with option to search omitted results as well, relevance based on sites' linkages & authority. Displays total number of hits found and the amount of time it took to complete the search, displays 10 hits per screen. With multiple results from the same Web site, the most relevant result is listed 1st with the other relevant pages from that site indented below it.	Home page is user friendly and not overly busy. Shows title, first two lines of page text, occasional "last refreshed" date, only one page per site appears in top results, "translate" option using Systan software, "more pages from this site", "company fact sheet" about the company that owns the site, "Related pages" link, sponsored matches from Overture, file size, 10 records at a time, offers related search suggestions, maximum of 1000 records can be displayed
15. Special features	Clustered results with "more results from..." option. Search any language, page specific search, some specialized searches, searches for the ampersand & and the underscore _ characters. Other features include calculator, definitions, "I'm feeling lucky", telephone book, Web page translation, street maps, and search by number (UPS tracking: FedEx tracking, Patent numbers, FAA airplane registration, FCC equipment IDs). Special services: Google Answers (open forum where researchers answer	Types of searches offered: Web, image, audio, video, directory, news & Web master. SE limits, offers simple and advanced search. Detects searcher's region of origin, option to select non-Roman alphabet entry, translation into nine languages

FEATURES:	SEARCH TOOL / ENGINE:	
	Google	**AltaVista**
	questions), Google Catalogues, Google Groups (Usenet discussions), image search, Google Labs, Google news, Google wireless. Supplemental Results feature.	
16. Documentation	Google Help Pages, Google Zeitgeist (search patterns & trends), Press Releases, copyright information, site map.	Search Help File, Special Syntax Help File, Press Releases, Business Services, About AltaVista, Submit a Site, copyright information.
17. Partner changes		
18. Partner size comparisons		
19. Size and scope	Leading, over 3 billion pages	3rd position, over 1 billion pages
20. Overlap	Receives from: Google AdWords Pay Service, ODP Directory, GoogleBot Spider, as does AskJeeves, Lycos, Search Netscape, AOL & Yahoo Database used by AOL, iWon, Netscape's Search, Backend SE at Yahoo, Weather Underground.	Receives from LookSmart Directory & Overture Paid Services, as does LookSmart, MSN, and Yahoo & Lycos. Little overlap with other SE.
21. Duplicate detection	Grouped under categories	Grouped under one title
22. Phrase search	Quotation marks	Some automatic, use quotation marks, use punctuation marks between words, e.g. hyphen – or comma
23. Spelling suggestions	Available "Did you mean…"	"Did you mean…"
24. Portal features		
25. Freshness	1 month average	3 months average, "refreshed in the past 24 / 48 hours" indicates a refreshed page
26. File types	Searches 12 file formats. PDF:, MSWord (.doc):, PowerPoint (.ppt):, Excel (.xls):, PostScript (.ps):, WordPerfect (.wpd):, .txt, .rtf, .asp. Opportunity to "View as HTML" to avoid viruses and downloading files.	Various, including PDF:
27. Family filters	Available	Available
28. URL search (find a single page)	allinurl: inurl:	url:
29. Site search (all URLs from a particular site)	Site: Combine with search term	Host: Add search term to narrow

FEATURES:	SEARCH TOOL / ENGINE:	
	Google	**AltaVista**
30. "New Classification"	Search Engine	Search Engine
31. INCONSISTENCIES		Time outs, limits, case sensitivity, diacritics, field searching

10.1.2 Yahoo and Lycos

Table 5: Yahoo and Lycos features

FEATURES:	SEARCH TOOL / ENGINE:	
	Yahoo	**Lycos**
1. Databases	Yahoo Directory, sponsored links (ads from Overture), Google for Web Pages (the SE), Images (Google), Yellow Pages, products, other databases provide much of he information from the portal side of Yahoo	Ads (sponsored links): Overture & Lycos' AdBuyer, Web results: Lycos Network Content, 10 LookSmart Ads, FAST (AlltheWeb) database, news: Lycos & FAST News, directory: Open Directory (by link), images: under multimedia, FAST database, Audio & video: under multimedia, FAST database
2. Strengths	Very popular, one of the best-known Web sites, one of the larger directory databases, many services for popular and general information. Cached option, "view as HTML" option, shows Yahoo category links if applicable.	Additional access point to the FAST database, some advanced features, extensive portal content, supports 47 languages, including Afrikaans
3. Weaknesses	Some content is dated, much emphasis on commerce, attempts to keep users on Yahoo Properties.	Web results section can be confusing, excessive ads, including the LookSmart ones on the top of the Web results, no full Boolean searching, missing advanced features, FAST database version sometimes older than AlltheWeb.
4. Default operation	Sometimes defaults to AND, sometimes to OR, uses its so-called "intelligent default", but Yahoo does not clarify its use – thus defaults vary.	AND
5. Boolean searching	Boolean & nesting are not supported, use + to require a term & - to exclude a term. Search text boxes in Advanced search provide options similar to AND, OR, " ", NOT.	Only - and + supported, drop down menu in advanced search with options like "all the words", "any of the words" etc.
6. Stemming		

FEATURES:	SEARCH TOOL / ENGINE:	
	Yahoo	**Lycos**
7. Proximity searching	Phrase searching (double quotes), this turns off the automatic truncation	Phrase, using double quotes
8. Truncation	Automatic, search terms of more than five or six characters are automatically truncated. A term with double quotes will not be truncated, the * is used to truncate a term of one to five characters. Internal truncation is supported, no truncation at the beginning or a term or in phrase searching	Not available
9. Case sensitivity	No	No
10. Field searching	Title:, url: or truncated as t: and u:	In advanced searching with drop down menus: title:, url:, link:…in the…Title, …in the…URL, …in the…Referring URL
11. Limits	Advanced search: limit to Yahoo Directory categories or Yahoo Directory sites. Time limits (options provided), language, site/domain from drop down menu	Language, domain, URL, site, 47 languages, language limit can not be combined with a filed search
12. Stop words	Directory search ignores some common stop words	No stop words are excluded, choose to search stop words
13. Sorting	By relevance. Shows results in six categories: Web (from SE – Google) in Google's relevance order, Images (Google), Directory: (Yahoo Directory), Yellow Pages (Yellow Pages search form), News (Yahoo News Database), Products (Yahoo Shopping Search).	By relevance, no options to sort by site or date or alphabetically
14. Display	The home page display is extremely busy. Directory results display site title, description, URL & category name. Ratings: "Most popular" is displayed at the top, then "pick", with a review, then "cool". Total number of hits is displayed, showing 20 Web results per page (customizable).	The home page layout is very busy. Results: Only ten hits at a time, showing title, key word in context extract & URL. No date, language or file size available. Web results display: 1st under Web Results: links from Lycos Network and ads, then ten listings from LookSmart, then follows FAST Web search database results. Displays also 2 - 4 "sponsored link" listings from Overture, rest of the e1st six listings are from Lycos' AdBuyer. The "Sidesearch" link opens up the result on the right & move search results to a frame on the lefty. Translation capability

FEATURES:	SEARCH TOOL / ENGINE:	
	Yahoo	**Lycos**
		using Systan software.
15. Special features	Local Yahoos: 10 countries in Europe, 10 countries in Asia Pacific, four countries in the Americas, with US in Chinese & Spanish, as well as customization features for US cities such as Atlanta, Boston, Chicago, LA, NYC etc. it also offers special sections on general guides, small business, enterprise & personal finance. Use ! to get around in Yahoo Quickly. Option to set search preferences.	Web search engine, subject directories
16. Documentation	Search Help, Yahoo Information, Press Releases, How to Suggest a Site, Company Information, Copyright Policy, Terms of Service, Jobs, Advertise with Us, Privacy Policy.	Lycos Search Documentation, Press Releases, Privacy Policy, terms and Conditions, Add Your Site to Lycos, Lycos Search for Missing children, Help, Feedback, Jobs, Advertise, Business Development, Enterprise Services.
17. Partner changes		
18. Partner size comparisons		
19. Size and scope	Guessed at over 3 million records in the directory	
20. Overlap	Receives from: Overture Paid Services, GoogleBot Spider as does AOL & AltaVista, MSN, Lycos, Google, and Search Netscape.	Receives from: ODP Directory, Overture Paid Services & FAST Spider, Lycos Ad Buyer Pay Services as does Google, Search Netscape, AlltheWeb, AskJeeves, MSN, AltaVista & Yahoo
21. Duplicate detection		
22. Phrase search	Yes, " "	Yes, " "
23. Spelling suggestions		
24. Portal features	Yes, extensive. Search preferences can be saved to one's account.	Extensive content
25. Freshness		
26. File types		
27. Family filters	Available	Available
28. URL search (find a single page)	u: Only Yahoo	
29. Site search (all URLs from a		

FEATURES:	SEARCH TOOL / ENGINE:	
	Yahoo	**Lycos**
particular site)		
30. "New Classification"	General Portal / Hub	Search engine
31. INCONSISTENCIES		

10.1.3 Excite

Table 6: Excite features

FEATURES:	**SEARCH TOOL / ENGINE:**
	Excite
	No longer a separate SE. Excite.com does not search its own database any more. It provides Overture paid positioning results & provides Inktomi results from Overture. The directory is now the Open Directory, while the news search uses Dogpile's meta news search.
1. Databases	Uses own directory database Excite Channels along with a current news database (News Tracker / Excite News) & several reference databases e.g. dictionary, almanac & encyclopedia. It has a small database of customized links to popular search topics. These are displayed 1st, followed by directory matches, & then Web page matches from Excite. News database & Reference database records may be displayed with separate headers. Meta SE searching Google, LookSmart, Inktomi, AskJeeves, About, Overture, FindWhat, FAST, Open Directory, Search Hippo & Sprinks. Excite Directory, Excite News Search, Excite Photo Search
2. Strengths	Personalization features & high relevance on popular topics. One of the smaller SE but very well known, it offers sophisticated personalization, excellent relevant results for very popular queries, and the News Search provides access to online versions of newspapers, magazines and news wires. Easy to follow search help section, particularly for the inexperienced searcher. Its headings and links are well organized. Uses InfoSpace meta search technology.
3. Weaknesses	Boolean operators must be in upper case, it has a smaller database, does not support truncation or field searching.
4. Default operation	Multiple search terms are processed as an OR operation
5. Boolean searching	AND, +, NOT, -, OR, AND NOT. Excite will only search SEs that support these features. Use parenthesis for nesting. Boolean operators must be in upper case: Boolean operators and search math are not supported in Advanced search. Select functions from drop down menus.

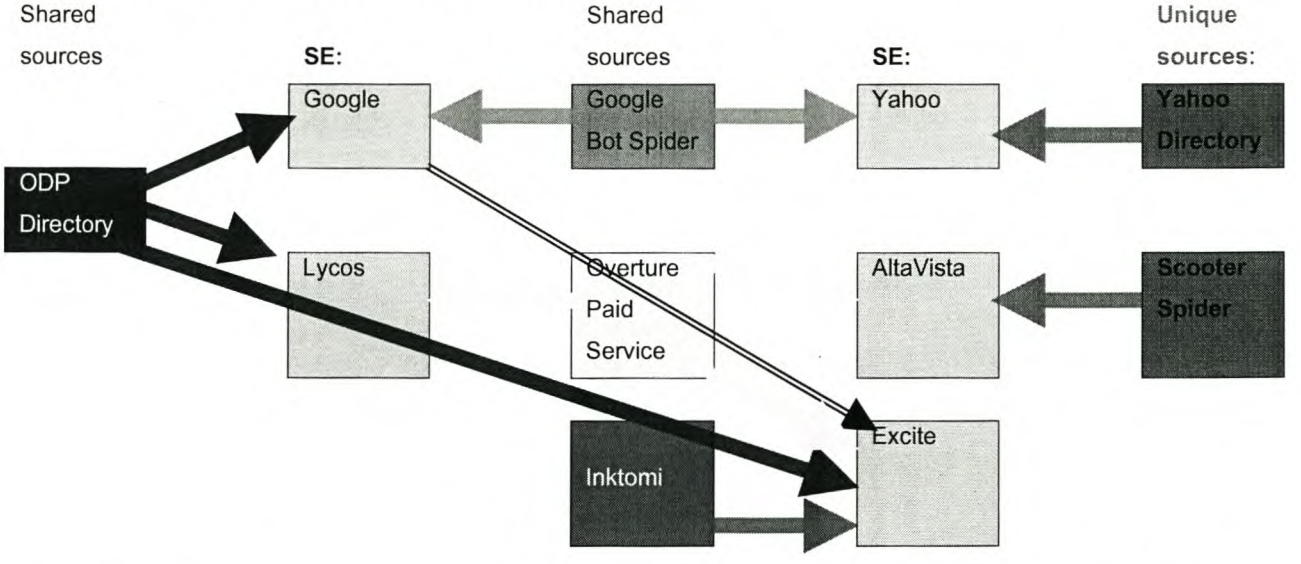
	SEARCH TOOL / ENGINE:
6. Stemming	
7. Proximity searching	Indicate a phrase search by using double quotes around a search phrase. Stop words in a phrase are ignored & will return result with and without requested (or other) stop words.
8. Truncation	Not supported
9. Case sensitivity	Not case sensitive, both upper & lower case results are shown.
10. Field searching	Not supported
11. Limits	Automatically limits to English language searches. No other limits available on regular search. More Search / Advanced Search gives the option to limit the search to any of a list of nine languages. Advanced Web Search can limit by country, common US top level domains & adult content filter.
12. Stop words	Words & numbers from the stop word list will not be searched.
13. Sorting	By relevance with groupings by site available at the end of each brief record. Every hit has a "more from this site" link regardless whether other sites are available or not. No options to sort alphabetically or by date.
14. Display	Includes the relevance score, title, URL & a brief summary. Provides the option to do a follow up search – use the "more links like this" link that creates a new search for records similar to the one chosen. Ten records are displayed at a time, except when results are sorted by site. Excite then displays the top 40, but will not provide an option for displaying the rest arranged by site. The Show Titles Only option displays 20 at a time, showing relevance score, title & "More Like This" option. Up to 50 records at a time can be requested in the Advanced Web Search.
15. Special features	Highly advanced personalization capabilities.
16. Documentation	Help Index, General Search Help, Advanced Search Help, and Press Releases.
17. Partner changes	No partners are currently using Excite.
18. Partner size comparisons	One of the smaller SE
19. Size and scope	
20. Overlap	Little overlap with other SE
21. Duplicate detection	
22. Phrase search	With " "
23. Spelling suggestions	
24. Portal features	Yes: email, personalize stocks, news, horoscope, weather, scores, etc.
25. Freshness	
26. File types	
27. Family filters	Only in Advanced Web Search.
28. URL search (find a single page)	Not listed but enabled url:
29. Site search (all URLs from a particular site)	site:

	SEARCH TOOL / ENGINE:
30. "New Classification"	Search Engine
31. INCONSISTENCIES	

10.2 Major search engines: overlap and shared sources

Figure 7 is based on the literature studied and was devised to indicate the major search engines (as identified for this study) and their overlap of sources together with any unique sources – i.e. sources that are not shared with any other major search engine partner.

Figure 7: Overlap and shared sources of the five major search engines



The diagram shows that Yahoo and AltaVista are, within the major search engines identified, the only search tools with unique sources. ODP is searched by Google, Lycos and Excite, but not by Yahoo and AltaVista. Yahoo and AltaVista have unique sources, and share Overture results. Yahoo obtains results also from the GoogleBot spider (sharing with Google), that AltaVista does not share in.

This sharing of resources is dynamic, and the prudent searcher has to familiarize himself periodically with these relationships as they influence search results. A searcher may want to "spread the net", as it were, as widely as possible, and to include search tools that add value (and not only "also cover") in a search. Searching on Google, AltaVista, Yahoo and Excite will also cover results from Lycos, and would limit possible duplications in the context of the major search engines identified.

11 HOW DOES SEARCH ENGINES SEARCH?

11.1 Search engine search approaches: initial search basis

Level Ten (2002d) distinguish between two search types – Web Page Search and Web Site Search, and further distinguish between two search scopes – External Search and Internal Search. Often popular search engines are only portals to someone else's database. Yahoo, for example, maintains its own directory for Web results yet uses Google for its Web page results. Yahoo used to use Inktomi for its Web page results, which is also used by HotBot and others.

11.1.1 Web Page Search

Indexes the content of Web pages to create a list of pages that best match a search phrase.

11.1.2 Web Site Search

Indexes text description for a site.

11.1.3 External search scope

Pages or sites across multiple domains, such as Yahoo and AltaVista, are searched. Most Web sites are found using popular external search engines. Thousands of external search engines are available for public use (Level Ten, 2002b).

11.1.4 Internal search scope

Used to search pages within a designated Web site. An internal search engine allows searching within a site for pages containing certain key words. Internal searches are particularly useful on large sites as an alternative to hierarchical menus for site navigation. Many scripts or programs are available to add searching capability to a site. Level Ten (2002a) estimates that over half of Web users are search dominated versus hierarchy / links dominated. Web sites should offer both options to users.

11.2 Search engine searching

Conceptual knowledge and the ability to understand the output as related to context is all that is required to successfully source information on the Internet. Searchers interact with the most Web-based programs in non-technical and normal conversation. The complexity of calculation exists in the software itself (Day, 2001). The magic happens at the user interface and search engine software. The search engine must accept the search query and compare this query to the items in the catalogue, where link analysis and relative page value are dynamically created. All search engines go about this differently, resulting in different results. The biggest difference between search engines is their ranking algorithm. To compensate for inaccuracies, search

engines try to keep their algorithms secret, thus protecting the quality of their results (Wall, [2003e]).

11.2.1 Different sources

Search engines typically return results from many different sources such as directories, pay services and indexes (Ayache, 2003). The results from one domain are grouped together (clustered) to prevent multiple pages from one site appearing many times in the results. Most search engines provide a 'more results from this site' link to allow viewing of these pages (Spannerworks, 2003b).

11.2.2 Relevance ranking

Relevance ranking is determined by the location and frequency of keywords and phrases in the web document or metatags as well as the number of hyperlinks that are pointing to the sites. The more links, the greater the popularity and value of the page (Zaino, [2003d], Spannerworks, 2003b). Savoy and Picard (2001: 564) suggest that hyperlinks provide very useful information for extracting patterns representing various cyber communities or sets of authoritative pages relative to broadly represented topics.

The study done by Gwizdka and Chignell ([1998]) points out that the relevance in an inter-linked collection of documents is not only determined by each document considered separately, but also by the inter-linked structure of the whole system. Given such an approach, even the "non-relevant" document may become partially relevant when linked to relevant documents.

In her study, Hirsh (1998, in Wang, Hawk & Tenopir, 2000: 231) found that a variety of factors are used to judge relevance: topicality, peer interest, novelty, recency and convenience.

The problem with relevance ranking is that relevance is such a complex item. It has many topical aspects, e.g. who is assigning the relevance and which relevance are to be considered (Ljosland, 1999). Relevance *is* relevant to the user, the search and the topic.

A process called collaborative or 'social filtering' could also be used to determine link popularity and from that, ranking. Collaborative filtering is the process of filtering documents by determining what documents other users with similar interests and/or needs found relevant (Weiss, 1997).

Ranking algorithms (the methodology by which search engines calculate positioning results) can be influenced by a variety of factors including domain name, spiderable content, submission practices, and HTML code and link popularity. Many search engine algorithms score the words

that appear toward the top of the documents more highly than the words appearing at the bottom. Words in HTML header tags (H1, H2, H3, etc.) are also given more weight by some search engines (Barlow, 2002a). Search engine ranking algorithms are closely guarded and constantly updated to attempt to filter out those sites that attempt to manipulate the results (Spannerworks, 2003b). Search engines frequently update and change their algorithms to ensure quality of results and to prevent people from figuring out exact key mathematical ratios on which searches are based (Wall, [2003d]).

In *Results and challenges in Web search evaluation* (1999: 1321), the authors argue that aspects of results ranking efficiency should include –

- Whether the Web pages returned to the user are relevant (precision),
- Whether the pages are presented in order of relevance,
- Whether a significant or desired number of available relevant pages have been identified to the user (recall),
- Whether a required fact has been found and presented,
- Whether a significant or desired number of aspects of the user's search needs have been covered by the set of pages returned,
- Whether returned pages are authoritative,
- Etc.

Results and challenges in Web search evaluation (1999: 1329) points out that ranking problems might be the result of particular spidering and crawling, and intimates that efficient spidering might lead to more efficient ranking of results.

Google's PageRank relies on the uniquely democratic nature of the Web. It uses its enormous link structure as an indicator of an individual page's value. Thus, Google interprets a link from Page A to Page B as a vote, by Page A, for Page B. Google does not only examine the volume of votes, or links, a page receives. It also analyzes the page that casts the vote. Votes by pages that are considered "important" weigh more and help make other pages more "important" – it bears testimony to their value. These important, high quality sites receive a higher PageRank. To ensure a match for the query, Google combines PageRank with sophisticated text matching techniques to find pages that are both important and relevant to the query (Google, 2003g).

In his fascinating article Rogers (2002) explains the Google PageRank algorithm and how it works. However, exploring the deeper aspects and relationships of Google's PageRank Algorithm falls outside the scope of this study.

11.2.3 Text searching

Barlow (2002a) lists two main methods of text searching: keyword searching and concept-based searching.

- *Keyword searching*

This is the most common form of text search on the Web - most search engines do their text query and retrieval using keywords. Unless the author of the Web document specifies the document's keywords (meta tags), it is up to the search engine to determine them. Search engines pull out and index words that are believed to be significant. Words towards the top of the document and words that are repeated several times are more likely to be deemed important. Some sites index every word on every page, while others index only part of the document. Lycos indexes the title, headings, subheadings, and hyperlinks to other sites along with the first 20 lines of text. Full text indexing systems generally pick up every word in the text (excluding stop words). AltaVista claims to index all words, including stop words (Barlow, 2002a).

- *Concept-based searching*

Unlike keyword search systems, concept-based search systems try to determine what the searcher mean, not just the 'what' of the search. Ideally, a concept-based search engine returns hits on documents that are 'about' the subject/theme being explored, even if the words in the document does not match the search terms exactly. The search engine Excite relies on concept-based searching, also known as clustering. This means that words are examined in relation to other words found nearby. There are various methods of building clustering systems, and they are highly complex and rely on sophisticated linguistic and artificial intelligence theory. Excite uses a numerical approach – the software determines meaning by calculating the frequency with which certain words appear. When several words or phrases that are tagged to signal a particular concept appear close to each other in a text, the search engine concludes, by statistical analysis, that the page is about a particular subject. Concept-based searching works best when many search terms that refer to the subject searched are used (Barlow, 2002a).

Wall ([2003a]) speculates that the information revolution will bring about software that will be more able to understand emotion and what the searcher is really 'searching' for. Computers will also become more understanding of speech and apply the appropriate words to the sounds searchers make. Dragon Naturally Speaking is already doing a good job of this.

Contextual hints to specific meaning are now often included in advanced applications of artificial intelligence to searching software. For example, if a searcher types in 'survivor' in the health section of an indexed search engine, they are probably not looking for information on the

television show 'Survivor'. Such linkages of topics within a specific domain based on context illustrate how computers are becoming 'intelligent'. Technology is able to filter out the irrelevant links and exclude sites that do not fit a certain profile. The software notes the placement and pattern of words and then sends a crawler software agent into the specified pages on the Web that matches the specific criteria. This strategy also works when the search software considers the searcher's last 'stop' on the Web. Chances are, if one is looking for health information, the previous URL in the browser was a health-related site. New software is able to use this context in its criteria for the search parameters (Day, 2001).

11.2.4 Browser window

On the search engine home page the search begins on a Web page displayed within the browser window. On the Web page, there is a field to enter the search topic. Most sites have a Help section or a Frequently Asked Questions (FAQ) section. Searchers could start with the simplest search option on the search engine, and then progress to more advanced search options if required. Day (2002) describes the following strategies for selecting search engine search options: simple search, advanced search - phrase search and Boolean search.

When a searcher enters a search term into search engine or directory the search tool then uses its algorithm so search its database of pages or sites to find a matching key phrase and return a list of results (Spannerworks, 2003b).

11.3 Searching aids for the searcher

Key word searching can be used for a specific word or phrase. Browsing in subject indexes is the best strategy when one is still trying to work out exactly what information is required. The best possible searching aid is probably the searcher's ability to familiarize himself with different search engines and their capabilities.

11.3.1 Simple search option

Simple search criteria are entered into an editable field on screen. The search is activated by pressing the search button. The specific search term can be broadened if the first batch of results are inadequate.

The simple search generally requires -

- Specificity
- Nouns as query words
- Main topic words first
- Six to eight key words
- Truncation with * to add plurals.

11.3.2 Advanced search options

Most search engines and directories allow for advanced searching, but methods may vary. Usually phrase and Boolean searches are supported. Many search engines allow for application-based rules within each process.

❒ *Phrase search option*

Allows the searcher to search on multiple words for one topic. Most search engines will do a phrase search without the searcher needing to inform the search engine that the words have to be grouped together. The software assumes that all the words have to be present on the page and in close proximity to one another.

❒ *Boolean search option*

Boolean searching refers to a form of logic applied to the search. A Boolean search requires additional words to be used. It also allows for excluding Web sites. Some Web searches allow the searcher to click on the parameters that must be employed, e.g. 'any' equals the Boolean ''OR'',
and ''all'' is similar to ''AND''. Boolean searching allows software to quickly narrow searches to enable results to pinpoint the required information. The addition of terms to the specificity of the search will mostly result in better results.

Boolean search parameter definitions:

- **'AND'**: search on Term1 AND Term2 – AND narrows the search by retrieving only documents that contain every one of the key words requested
- **'OR'**: search on Term1 OR Term2 – OR expands the search by returning documents in which either or both keywords appear
- **'NOT'**: search on Term1 but NOT Term2 – NOT or AND NOT limits the search by returning only documents containing the first key word but not the second
- **(Nesting)**: (Term1 OR Term2) AND Term3 - using parenthesis to combine several search statements, also used to separate keywords when more than one operator is used. Always enclose OR statement in parenthesis

Boolean logic is not supported by all search engines.

- **Proximity operators** (NEAR, SAME, ADJ or FOLLOWED BY) are not part of Boolean logic, but have a similar function in formulating search statements: Term1 NEAR Term2.
- **Stop words**: some search engines ignore short, common words that may appear in the text or title of documents: a, an, and, as, at, be, if, into, it, of, the, to, on, or. Stop words have

little or no semantic content, and may have a high frequency across a collection (Weiss, 1997). Most search engines ignore stop words, but Google allows stop word searching by using the plus (+) sign.

- **Implied Boolean operators ('search math'): +, -, " ", ***

Implied Boolean operators (Zaino, [2003b]) or search math implies using the plus (+) and minus (-) symbols directly in front of terms to force their inclusion and/or exclusion in a search: +Term1 –Term2. Double quotation marks (" ") are used around phrases to force the search engine to search in that EXACT word order: "Term1 Term2". Quotation marks should not be placed around a single word. The asterisk (*) can be used as a truncation device to search variations in spelling and word form, e.g. library* will return library, libraries, librarian, and librarianship. Some search engines support end truncation (college* > college, colleges, collegium, collegial) as well as internal truncation (col*r > colour, color, colander). Phrases with key words can be combined by using the double quotes and the (+) and (-) signs, e.g. +"fish river canyon" +hiking. Stemming can also used to broaden a search. Stemming is related to truncation, and refers to the ability of the search engine to find word variants such as plurals, singular forms, past tense, present tense, etc. Stemming is the process of removing prefixes and suffixes from words in a document in the formation of terms in the system's internal model to group words that have the same conceptual meaning, e.g. WALK, WALKED, WALKER and WALKING (Weiss, 1997). Some stemming covers only plural and singular forms.

Keywords and phrases typed in lower case retrieve both lower and uppercase versions.

Results for a search request may vary depending on the site selected and the time of search. The Web is so dynamic that the results from the same site can vary by the minute. This vigor adds the advantage of instantly distributing knowledge that is time and context dependent. The relevance of each response is determined by the user (Day, 2001).

12 INFORMATION RETRIEVAL

Information cannot be retrieved if it cannot be found. For information to be found, it must be organized and described in an organized, agreed upon way that enables different people to describe and organize the information in a similar way. Information organization is a fascinating and complex aspect of information retrieval, and several systems have been developed to organize information. This study will not venture further into the field of information organization but to mention that information is usually described in terms of some or all of the following ways:

- Source
- Language
- Identifier
- Description
- Type
- Format
- Coverage
- Date
- Relation
- Contributor
- Subject
- Title
- Creator
- Publisher, and
- Rights.

Optimal information retrieval can be defined as: Find all the relevant and none of the irrelevant documents. Jansen (1996) discusses three major information retrieval paradigms: statistical, semantic and contextual.

- Statistical approach: emphasizes statistical correlations of word counts in documents and document collections, using schemes such as vector space models for document representation and retrieval. Capturing term associations in documents is another example.
- Semantic approach: views documents and queries as representing some underlying meaning. It emphasis natural language processing or the use of artificial intelligence queries.
- Contextual approach: takes advantage of the structural and contextual information typically available in retrieval systems, e.g. taking advantage of thesauri and encoded relationships among terms.

There are two accepted standards of performance for comparing and evaluating retrieval systems in the field of Information Retrieval: recall and precision (Jansen, 1996, Gwizdka & Chignell, [1998]):

- Recall = Relevant Documents Retrieved/Total Number of Relevant Documents
- Precision = Relevant Documents Retrieved/Total Number of Retrieved Documents

In an ideal world, both precision and recall is 100%, but this is rarely true in reality. Information retrieval systems therefore attempt to maximize both recall and precision simultaneously (Weiss, 1997).

Gwizdka and Chignell ([1998]) distinguished four different precision measures:

- Full precision – takes fully into account the subjective score assigned to each hit
- Best precision – takes into account only the most relevant hits
- Useful precision – takes into account only the most relevant hits and hits containing links to the most relevant ones
- Objective precision – objective because it does not rely on human relevance judgment, based on computed presence or absence of required terms and on the distinction between good and bad links.

The interconnected character of the Web and its expansive user population are major factors that are affecting the evaluation of information retrieval from the Web (Gwizdka and Chignell, [1998]).

Some authors studied prefer to omit **recall** as evaluation criterion of information retrieval because it is almost impossible to assume how many relevant items are available for a particular query on the dynamic Internet environment. Response time rather than recall is then used to evaluate search engine performance (Chu & Rosenthal, 1996).

Ljosland (1999) suggests that researchers publishing works on comparison of Web search engines use precision as their main evaluation measure, evaluating only the highest ranked hits. If the n first documents are evaluated, the precision found from these documents is called P@n (precision at n). Some researchers also use pooled recall (computed from the n highest ranked), and some evaluate qualitative properties like user interface and ease of use. These qualitative properties are also used in the evaluation tool designed in this study.

When selecting an appropriate search engine the searcher should therefore consider which information retrieval paradigm would most suit a particular search, as well as weigh the

Recall/Precision ratio and response time of a search engine. Applying field searching will generally enable the searcher to retrieve information that is described or organized in a particular way such as format, title, language, publisher, etc.

The general user using search engines may prefer a less technical approach, such as to what degree the system's (search engine) results support the larger process of information use. In other words – does the search results meet the search query? Information retrieval is usually part of a bigger process of information use, and returned hits, in the case of online search engines, must ultimately meet the information requirements of the searcher. The searcher has the responsibility to select the appropriate search tool for the specific query.

Savoy and Picard (2001: 564) point out that most web users prefer a high precision value and will readily accept a lower recall value when the search query is answered quickly, and when the search engine used can be deemed intelligent.

In-depth explorations of the "back-end" search engine algorithms and search software programs fall outside the scope of this study. This study focuses on the "search front" of search engines – what is presented to the user that wants to search for information using a search engine. To this end, search engine algorithms and search engine information retrieval paradigms will not be explored further.

13 USER-WEB INTERACTIONS

The typical user population searching the Internet are users who have very little if any training on how to conduct information searches, as opposed to trained specialists in information retrieval. Thus user interaction with the system is critical, because experts are generally better at adapting to different types of interface than are novices, whose performance is greatly affected by the type of interface used (Gwizdka & Chignell, [1998]).

Users are as heterogeneous as the resources they seek to use. The majority users are perpetual novices with diverse subject backgrounds and varying levels of information, computer, and Web literacy (Wang, Hawk & Tenopir, 2000: 230). Ideally, user differences and varying capabilities must be considered when designing a search interface for search engines. Desired characteristics of user interfaces are: easy to understand, easy to learn, error tolerant, flexible and adaptable, appropriate and effective for the task. *Methodologies and website development: a survey of practice* (2002: 382) includes navigation, function and graphics in the list of interface characteristics. Interfaces that are truly useful for all are referred to as every-citizen interfaces - ECIs (National Research Council, 1996, in Wang, Hawk & Tenopir, 2000: 230).

Jansen, Spink & Saracevic (2000, in White, Jose & Ruthven, 2003: 708) report that users tend to refrain from using the advanced search facilities offered by search engines. To accommodate such users, the user interface of search engines should incorporate functionalities that will help users and searchers search more effectively.

A summarization system specifically designed for web search engines can provide users with a means to effectively assess document relevance without referring to the full text of a web document. Document abstracts are often found to be out of date, ambiguous and too short. A summary biased to the user's information need (i.e. the query) can be beneficial (White, Jose & Ruthven, 2003: 709). The authors conclude that automatically generated web page summaries allow users to gauge document relevance more effectively than those presented by the traditional ranked title/abstract approaches (White, Jose & Ruthven, 2003: 729).

Individual user differences might affect how users search for information on the Web, and should therefore be considered when designing a tool for search engine evaluation.

Wang, Hawk and Tenopir (2000:231) refer to a number of studies undertaken to take user oriented approaches such as sense-making and cognitive and behavioral approaches into

account when investigating the complex manner of users' information retrieval actions. The affective and sensorimotor domains for novice users complement the cognitive elements of online searching. Other domains identified are the user and computer domains. The user domain consists of the following elements: situational (task), affective (intent), cognitive (knowledge structure) and query (characteristics). The computer domain consists of interface, engineering (hardware), processing (software) and content (Spink & Saracevic, 1998, in Wang, Hawk & Tenopir, 2000: 231).

Bishop and Starr (1996, in Wang, Hawk & Tenopir, 2000: 232) concluded that information systems designers for the increasingly heterogeneous user population with diverse sets of information needs must understand which aspects of searching behavior are universal and which are situation-specific.

Wang, Hawk and Tenopir (2000: 232) suggest a multidimensional model of user-Web interaction. Their approach is holistic, because not all users will search the Web in the same way. Individual differences may cause difficulties in finding appropriate information.

The model consists of three components: the user, the interface and the Web space:

- User
    - The user is influenced by dynamic situational factors (particular task, information need and user's knowledge state)
    - Individual characteristics play role (cognitive style, affective state before and during the search)
    - Cognitive behavior includes thoughts, search strategies, problem solving, decisions and mental models
    - The affective state might change as a result of the interaction
    - Physical factors are sensorimotor skills (hand-eye coordination and control of input devices)
- Interface - enables certain actions or interactions, and is deemed intelligent if it can provide context-sensitive help, such as response messages or suggesting a next step.
    - Access methods
    - Navigation tools
    - Access results/objects
    - Messages/clues
    - Input/output devices
- The Web consists of:
    - Objects, that provide content, expression, relation, structure and hyperlinks

- Activated objects (portions of the Web that have been activated during the interaction)
- Web spaces – a collection of networked objects accessible by a method
- Organization schemes
- Metadata.

Wang, Hawk and Tenopir (2000: 249) suggest that their proposed model can provide a clear framework for developing user instructions – important information for search engine designers.

Aspects of Wang, Hawk and Tenopir's model of multidimensional web user interactions will be incorporated in the design of the search tool in this study. The author believes that it is critical to consider the user and the user's capabilities and background when selecting a Web search tool.

Hölscher and Strube (2000: 345) identify and differentiate two types of user expertise: technical Web expertise and domain-specific background knowledge. Searchers that are able to apply both types of expertise were found most successful in their searches. Deficits in one or the other type of expertise led to compensatory behavior, e.g. domain-expert/Web-novices rely on terminology and avoid query formatting. Users with lower levels of knowledge are less flexible in their search strategies and return to previous stages of their search more often rather than trying new approaches such as changing the search engine. This behavior emphasizes the importance of the user interface, navigation tools and ease of use. Both kinds of user expertise aspects are incorporated in the evaluation tool developed in this study.

This study differentiates between the terms **users** and **searchers** – **users** indicates the general public occasionally using the search engine to find information, and **searchers** refers to professional users of the Internet, e.g. academics, researchers, information brokers, etc.

14 CONDUCTING AN ONLINE SEARCH

Searching for information on the Internet requires proper planning – planning a search strategy as well as deciding which search tool to use to execute the strategy.

14.1 Search strategy

Planning a search strategy involves considering details such as –

- Using synonyms for the search term or phrase
- Application of search math, Boolean operators and parameters
- Spelling variations
- Using broader and/or narrower terms and
- Use search phrases (nesting), etc.

14.1.1 Smart searching

Barlow (2002d) suggests the following principles of smart searching:

- Know where to look first. Many databases contain specific information that might be more useful than using a general search engine.
- Fine-tune keywords. When searching for nouns, keep in mind that most nouns are subsets of other nouns – search with the smallest possible subset that describes the search query, and be very specific. E.g.: rather search with ''Nissan'' than ''car''.
- Be refined. Always read the help files and use the available search refining options. Use phrases if possible, and apply Boolean operators and other parameters such as the truncation * (* represents 0-5 extra letters, use ** for unlimited letters) and search math. Barlow recommends that the search query be submitted several times, each time adding further refinements to narrow down the list of hit fit. Also, use the Boolean NOT to deny irrelevant hits.
- Query by example. Many search engines offer the option of ''query by example'' or ''find similar sites'' to the ones that come up on the initial hit list. This process is also known as relevance feedback.

Notess (2002c) also suggests -

- Search for the source of the information, or guess at it, especially where currency is important
- Diacritics (ñ, ë, è, ê, etc.) - in most search engines searching without diacritics matches most hits
- Use selected field searching where supported (Notess, 2003n), e.g.
  - title:

- intitle:
- url:
- inurl:
- sites:
- link:
- anchor:
- image:

- Use selected limits (on advanced search forms) where supported (Notess, 2003m), e.g.
  - Language:
  - Date:
  - File Type:
  - Media Type:
  - Page Size:
  - IP Range:

14.1.2 Specialized databases

Specialized databases (such as Newsgroups and Company information) offer another way to find information on the Internet, and are useful when one knows exactly what one is looking for.

14.2 Selecting a search tool

Choosing a search tool that suits the query is vital to find accurate, reliable and, if required, current information. The 'hit fit' depends on the search strategy and the chosen search tool.

The best search tool/engine depends on what type of information one is trying to source. Search engines/tools can be divided into beginner, intermediate and advanced categories. Beginner search engines have a larger database of material and a less sophisticated indexing mechanism. For general interest topics, a beginner search engine (e.g. Yahoo) is a good choice. For more detailed information a more sophisticated searching mechanism is required, e.g. Excite, Lycos, or AltaVista (FAQ's, 2002).

14.2.1 Sophistication criteria for search engines

The following are sophistication criteria for search engines (Search engine types, 2002):

- Indexing mechanism – what mechanism is used, and how selective and effective is it?
- Database size – too small a database may lack important components
- Quality control – dead links, expired pages, age of documents presented
- Indexing policy, process, and frequency.

There are three methods used in the indexing of a web site database (Habib & Balliot, 2003):

- Full text index: a database index that includes all terms and URLs, compiled by spiders. In reality, search tools use a filter to remove words that are considered unnecessary or impractical to search.
- Keyword index: a database index that is based on the location and frequency of words and phrases. If a name or term is mentioned only once or twice on the site, it may not be included in its index. Keyword searching is the most used and fastest growing indexing method. Compiled by spiders.
- Person index: an index created by individuals who review Web sites and select the most appropriate words and phrases to describe their content. It provides a directory that is high in relevance and is based on similar cataloguing methods used by libraries. Unlike the other two indexing methods, the human index adds the value of being reviewed.

Apart from the list above, the following aspects must also be considered when selecting a search tool/engine:

- User interface: page layout, how "busy" or cluttered the page is, the amount of advertisers and the prominence of paid listings. The purpose of the web page must be clear, and the search capabilities and options should be clearly visible. This includes guidelines on how to use the search engine, frequently asked questions, help files, as well as contact details of the owners and the web master. It should be clear when last the site/page was updated.
- Policy on Spam: are there so many pop-up pages of advertisers that it is virtually impossible to actually use the search feature?
- Privacy policy: will the search engine create a cookie on the searcher's activities? Will the searcher's email address and any other personal information be made available to others?

Wang, Hawk and Tenopir (2000: 248) point out that many users and searchers will not participate in any user instruction programs. In this light, browser interface and easy to use search options become critical success factors for online information retrieval.

Search engines should, most of all, deliver what they promise in terms of information retrieval.

15 SEARCH ENGINES – PROS AND CONS

15.1 Search engine trouble shooting

Useful as search engines are, they can also be extremely frustrating to work with, and complaints such as returning too much information, returning hits where the relevant information in the returned documents is hard to find, returning outdated and inaccurate or irrelevant, no logical organization of results, too many graphics that slow the search down and the lack of contact details of webmasters are common (Top 10 complaints, 2002).

Common resolutions (2002) suggests that when a searcher struggles to find information, he/she should try to select a different search engine, try different search terms (also terms used globally) and consider simplifying the search string.

More search engine trouble shooting tips:
Troubleshooting (Zaino, [2003j]):

- Search returns a '404 – file not found' message: the requested file has been moved, removed or renamed - use Google's cached feature to retrieve a copy of the document.
- Search returns a 'Server does not have a DNS Entry': the browser is unable to locate the server or host computer, the network may be busy or the server has been removed or taken down for maintenance.
- Search returns a 'Server Error' or 'Server is Busy' message: the server that is being contacted is offline, has crashed or is busy. Generally, one can try locating it again later.

Habib and Balliot (2003):

- Query does not have a counterpart in the search engine's index: searcher might not be familiar with the search engine's composing criteria. The searcher should study the search engine's help section and recompose the query.
- No matching information retrieved: the search engine did not index significant keywords while spidering the Internet because it employs abbreviated rather than full-word spidering in creating and maintaining its database: use a different search engine that uses full word spidering.
- Returned very irrelevant results: the search engine filtered out important key words in the query. One could try to use a search engine with a moderate sized database. Alternatively, one could use a subject search tool having a large database, such as Yahoo.

Search engines are helpful tools to access a major portion of publicly available pages on the Web, they are the best available tools for searching the Web, and they have large databases

(Zaino, [2003e]). Unfortunately, they often return irrelevant results and the “Invisible Web” is not indexed.

16 SEARCH ENGINE EVALUATION – INSTRUMENT DESIGN

16.1 Instrument for evaluation/selection of search tools/engines

The previous chapters investigated search tool/engine characteristics and desirable features in an attempt to answer the research questions posed. Much information was gained from the search engine analysis and comparison, and this information will be used to inform the search tool design in this chapter.

Chu and Rosenthal (1996) suggest an evaluation model that has five main elements:

- Composition of Web indexes: how the robots/spiders index, e.g. coverage, update frequency and portions of web pages indexed (title and first few lines of text or entire web page)
- Search capability: fundamental search facilities (Boolean logic, truncation, limits)
- Retrieval performance: precision, recall and response time
- Output option: number of output options and actual content of output
- User effort: documentation and interface (comfort, ease of use and understanding).

Chu and Rosenthal's evaluation methodology corresponds with the evaluation tool developed in this study. However, the author feels that a clearer distinction between the different role players in online searching would be beneficial to the users of the evaluation tool. From the information gained in the literature studied, it was decided to base the evaluation tool on the 'three major role players' in conducting a web based search using a search tool/engine: the search tool, the user and the search query.

Gwizdka and Chignell ([1998]) support the methodology outlined by Chu and Rosenthal above, but add the impact of interlinking to the list of aspects to consider as pointed out by Ding and Marchionini in 1996 (Gwizdka & Chignell [1998]).

Randolph Hock (as reported in Corman, 2002:83) suggests that factors to consider when evaluating a database should include the information covered, the indexing program and retrieval algorithm used, as well as the interfaces and portals available. Search engine users must consider how the search engine is constructed, why it works the way it does, and how it can be fully utilized

Clearly, evaluation or selection of an appropriate Web search engine / search tool cannot be done in isolation. The search tool is relative to the searcher and the search query, and these factors influence each other.

The searcher's proficiency at Web searching as well as his experience largely determine the kind of search tool that would be most suitable. Is the searcher a professional searcher that has much experience searching for information? Is the searcher proficient in searching for information on-line? An experienced searcher will most likely be less affected by a "user-unfriendly" search interface, and will have more initiative to explore the possibilities of the search tool. The experienced searcher will transfer search skills and abilities and will be able to find the required information, even with an unorthodox search tool. An inexperienced searcher, however, may require more from the search tool selected, specifically the browser window. The help files and search options should be very user-friendly to guide the user through the search process, and to provide search options and alternative search avenues when necessary. Search options should be clearly explained, and should not be too intimidating, providing for basic and advanced searches. The experienced searcher and the inexperienced user are likely to use different search tools, and have different search strategies and search approaches. How much the searcher already knows about the topic of the query is also important for selecting the right search tool.

The type of information required is also important when selecting a search tool. The best possible online source for the particular information must be selected, e.g. a specialized directory, a hybrid search engine, a portal / hub with customizable features or a general, natural language search tool. Also, consider which indexing method would be most appropriate to find the information. For general inquiries to discover more of about a topic, a directory might be appropriate. Very specific information searches might return better results using a search engine or even a Meta search engine. Intelligent agents can be of enormous help in finding, customizing and updating online information.

To obtain a good fit with the searcher, query and search tool the ranking algorithm and the display of the search results can help or hinder the search and searcher.

Search engine relationships and overlap of sources will influence the selection of a search tool. Many search tools share databases and sources – using different search engines that all use the same sources will not provide the "cast a wider net" effect desired. When selecting the search tools for the query, a variety of databases and sources will cover different areas of the Internet, and would be more likely to provide a variety of information.

Some search engines share their results with partners and one could essentially be searching the same databases but with different fronts. To avoid unnecessary overlaps users should select search engines with different or a variety of sources and that offer unique sources, e.g.:

- About.com includes Sprinkle pay service, Teoma spider and Inktomi spider
- AltaVista includes Scooter spider, LookSmart directory and Overture paid service
- Yahoo Includes Yahoo Directory, Overture paid service, and GoogleBot
- Google includes ODP, Google AdWords, and GoogleBot.

Database size and speed may be factors to consider when selecting a search tool, depending on the kind of information that is required. Google has the largest database and is currently the fastest. However, a smaller but very precise search engine might be better suited for a particular query. The freshness of a database might also be important to a particular search.

The search engine evaluation tool was designed to provide for a "balanced score/evaluation" approach to selecting an appropriate search tool. All three of the main aspects are closely linked and each influences the other to a greater or lesser extent, depending on the nature of the query, the searcher's proficiency and available search tools.

This search engine evaluation instrument serves as a benchmark and standard for web search engines and search tools in general. It points to the critical factors and features that a good search tool should feature.

Table 7: Search engine evaluation tool

SEARCH TOOL	**USER**
Absence of inconsistencies	Filters available
Variety of languages available	Special user requirements, e.g. user assistance.
Documentation: copyright information, Spam policy, privacy policy and protection	Searching: user friendliness and ease of use
Classification: search engine, directory, portal/hub, intelligent agent.	Search planning and strategy, e.g. availability of online tutorials and search strategy examples
Default operation	Search options required: keyword / subject
Display: results ranking, # of hits, text from hits, customization, "More links like these…", query biased summary	Search customization features
Freshness: spidering frequency	Portal / hub features offered: news, email, customization
Duplicate detection	Protection: privacy and Spam

Field searching: URL, title, domain, etc.	Search speed
Family filters availability	Pop-up ads
Case sensitivity	Knowledge of field of search
Boolean searching	Browser window layout, user friendliness, interface
Databases used/shared	General search proficiency, characteristics, cognitive behavior, sensorimotor skills
Cached pages available	Internet search experience
Variety of file types	Speed of downloads – graphics, text
Limits	Relate user to best suited search tool: search engine, directory, portal, intelligent agent
Overlap: database and source sharing	**COMMENT:**
Ownership – points to possible affiliations and independence	**QUERY**
Partner changes	Alternative spelling
Partner size comparisons	Authority of the sources, e.g. academic, government, non-government, commercial, personal.
Phrase search	Database freshness required?
Portal features – e.g. customization abilities	File types most likely to contain information
Proximity searching	Filters, family or adult.
Consistent quality control: moved pages, dead links, recall and precision	Format: text, image, sound, video, presentations
Ranking algorithm employed Paid services, keyword frequency, links frequency, "authority sites"	Information required: current and up to date
Sufficient search help for different users	Information required: general and popular
Search options offered Basic / advanced	Information required: very specific and exact matches
Site search (all URLs from a particular site)	Overall general, broad field of query
Size, speed and scope	Query: related fields / subjects
Sorting: clustering, customization features	Relate query to best-suited search tool: database size, meta search engine, search engine, directory, specialized directory etc.
Special features	Stemming and truncation possibilities
Spelling suggestions	Synonyms
Stemming	**COMMENT:**
Stop words	
Strengths	
Translation features	

Truncation	
URL search (find a single page)	
Weaknesses	
COMMENT:	

Figure 8 depicts the ideal fit between searcher, search query, and search tool:

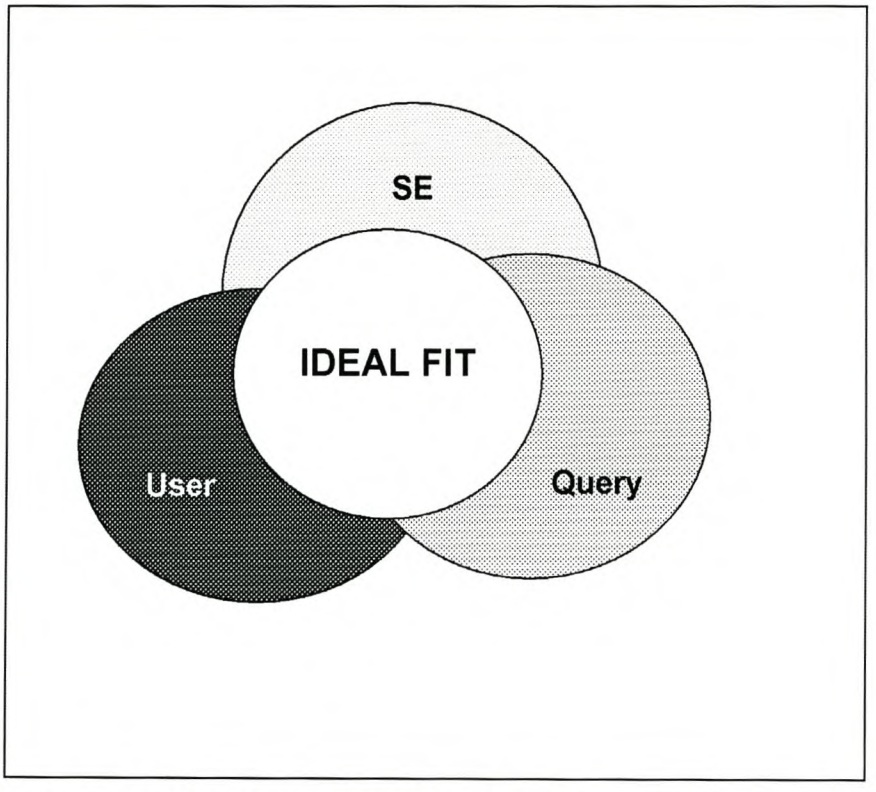


Figure 8: Searcher, query and search tool fit

There is no "perfect search tool", no panacea for all Web information queries. Optimum results can be achieved when one uses a combination of suitable search tools. The best search tool is one that is appropriate in terms of the searcher, the query, the format of results displayed and the type of search tool. It is proposed that the instrument developed to evaluate search tools will provide a balanced approach to evaluating search tools, considering all the factors that come into play.

17 LIMITED TEST RUN

The instrument designed for search tool evaluation was applied to three South African search engines to get an idea of how they would compare with the benchmark set by major search engines.

The local search engines were selected following informal discussions with professionals as to which local search engines they favor. Most responded that they use local search engines to search for local information. Ananzi and Aardvark were popular choices, and the author added Aha to the list.

17.1 Searching with local search tools

The author conducted two limited searches on three local search engines and briefly compared their search results. The same searches were also done on Yahoo and Google to set a benchmark for search results. The search terms used were *hertzoggies*, a South African delicacy, and *accommodation in Bothaville*. The search terms were not limited by country in the major search engines.

17.1.1 Test run parameters

No search operators were used in the first test run, except that the stop word "in" in the last term was omitted. The search was done on both the directories and web search facilities of all the search tools. In a subsequent search, search operators were used for the web searches in all the search engines, and the differences, if any, was noted. The international search options of the local search tools were not used because they all have international search partners, and the aim was to discover how they would search their own and local databases. *Bothaville* was capitalized throughout the search, and lowercase was used with *hertzoggies*. All searches were done using only the basic search facilities.

The search results are displayed in the tables below. The number of hits is shown, and the number of relevant hits in the first five hits were investigated. The results of the application of search operators are indicated. Only the first five hits were considered in terms of results overlap.

The searches were conducted on 15 January 2004 between 9:00 and 12:00 on http://www.yahoo.com, http://www.google.com, http://www.aardvark.co.za, http://www.aha.co.za, http://www.ananzi.co.za.

17.1.2 Yahoo and Google

Table 8: Yahoo and Google search results

SEARCH	Yahoo DIRECTORY	Yahoo WEB	Google DIRECTORY	Google WEB
Hertzoggies				
# Hits	0	15	0	51
Precision of 1st 5 hits	0	4	0	5
Comment		Search operator application: no change.		Search operator application: 23 hits.
Accommodation Bothaville				
# Hits	0	341	2	508
Precision of 1st 5 hits	0	4	2	5
Comment		Search operator application: no change		Search operator application: 231 hits.

17.1.3 Aardvark and Ananzi

Table 9: Aardvark and Ananzi search results

SEARCH	Aardvark DIRECTORY	Aardvark AFRICAN WEB	Ananzi DIRECTORY	Ananzi SA WEB
Hertzoggies				
# Hits	0	6	0	2
Precision of 1st 5 hits	0	5	0	Both hits were relevant
Comment		Search operator application: no change.	Very slow. Search operator application: no change.	Search operator application: no change.

SEARCH	Aardvark DIRECTORY	AFRICAN WEB	Ananzi DIRECTORY	SA WEB
Accommodation Bothaville				
# Hits	0	66	1868	45556
Precision of 1st 5 hits	0	5	1	1
Comment		Search operator application: no change.		Search operator application: no hits were found

17.1.4 Aha

Table 10: Aha search results

SEARCH	Aha DIRECTORY	WEB
Hertzoggies		
# Hits	0	13
Precision of 1st 5 hits	0	5
Comment		Slow. Search operator application: no change.
Accommodation Bothaville		
# Hits	105	0
Precision of 1st 5 hits	3	0
Comment		Search operator application: no change.

17.2 Applying the search tool evaluation instrument

The Ananzi, Aardvark and Aha features were investigated using the search tool evaluation instrument as benchmark.

17.2.1 Aha

Web address: http://www.aha.co.za

Table 11: Search tool evaluation: Aha

Aha	
SEARCH TOOL	**USER**
Inconsistencies	Filters available
Use of search operators made no difference in the search results.	*"Automatic – "offensive, obscene and rude"*
Languages available	Special user requirements
See partners.	*E.g., ease of use and guidance offered.*
Documentation: copyright information, Spam policy, privacy policy and protection	Search user friendliness and ease of use
Privacy and Spam policy	*Very easy to use, site map explains the layout and what each section searches.*
Classification	Search planning and strategy
Directory, offers SE capabilities for searching the WWW.	*Does the SE offer a search tutorial with examples?*
Default operation	Search options required: keyword / subject
See partners	
Display: results, # of hits, text from hits, customization, "More links like these…", query biased summary	Search customization features
Layout is user friendly, hits are not numbered. Displays total number of hits, nine per page, title, one or two sentences from the text and URL.	*Can customized parameters be saved? See partners*
Freshness: spidering frequency	Portal / hub features offered: news, email, customization
See partners	*News and travel Guide.*
Duplicate detection	Protection: privacy and Spam
No duplicates in search results noticed in test run.	*Documented*
Field searching: URL, title, domain, etc.	Search speed
Local domains, See partners	*Too slow a SE leads to user frustration.*
Family filters	Pop-up ads
Applied to the Global Directory that is based on ODP "offensive, harmful and obscene" sites were	*None, although there are banner ads, these are less intrusive than pop-ups.*

Aha	
removed.	
Case sensitivity	Knowledge of field of search
See partners	*Individual*
Boolean searching	Browser window layout, user friendliness, interface
See partners. Search operators are not indicated for local searches.	*Easy to find important links. Layout is user friendly, hits are not numbered but the total number of hits is indicated. Nine hits are displayed per page.*
Databases and searches	General search proficiency, characteristics, cognitive behavior, sensorimotor skills
Google, AlltheWeb, DMOZ. *RSA Directory: searches the database of SA web sites listed in the RSA Internet Directory. Only the contents of the home page, including the "description" and "keyword" Metatags are searched.* *The SA Web: searches the .co.za, .ac.za and .org.za domains. Results may include hits from the WWW, depending on the keywords used. Results are enhanced by DMOZ and Google.* *The Global Web: searches the WWW, results enhanced by DMOZ and Google. News: Moreover.com.* *Aha retrieved the least hits of the three local search tools, with the least relevancy.*	*Searcher or user, professional or amateur searcher – it is very individual*
Cached pages	Internet search experience
See partners	*Browsing, surfing, searching experience?*
File types	Speed of downloads – graphics, text
See partners	*Fast, no graphics on home page to slow it down.*
Limits	Relate user to best suited search tool: search engine, directory, portal, intelligent agent
Local domains ending in .za.	***Access to special features?***
Overlap: database and source sharing	**COMMENT: very user-friendly tool that makes searching a breeze, the Partners inspire confidence in the search results.**
Google and ODP (DMOZ).	**QUERY**
Ownership – points to possible affiliations and independence	Alternative spelling
The Aha Internet Directory	*A spell check would be useful. Searcher should*

Aha	
	consider spelling variations, stemming and truncation.
Partner changes	Authority of the sources
Dynamic.	*E.g. academic, personal web page, government site, commercial or not for profit organization.*
Partner size comparisons	Database freshness required?
Google and ODP are leading SE and Directories	
Phrase search	File types most likely to contain information
See partners	*Sound, text or images?*
Portal features	Filters
No	*Applied automatic filter might skew results.*
Proximity searching	Format: required: text, image, sound, video, presentations
See partners	*Will additional software be required to download the information?*
Quality control: moved pages, dead links, recall and precision	Information required: current and up to date
Links to the Global Directory (based on ODP) are regularly updated and some categories have been modified. *Retrieved least hits with lower precision during the test run.*	*E.g. latest news breaks, latest research.*
Ranking algorithm Paid services, keyword frequency, links frequency, "authority sites"	Information required: general and popular
Sponsored links displayed on the right hand side of directory categories. See partners.	*Easy to find almost anywhere, common topics or searches.*
Search help	Information required: very specific and exact matches
Link is easily spotted but does not supply and provides adequate assistance, and an overview of international search tools is given instead. Contact details of web master are provided.	*E.g. academic information.*
Search options offered Basic / advanced	Overall general, broad field of query
RSA Directory, The South African Web, The Global Web. *Search and Fart Search (Displays titles only)*	*As can be found in a directory or library gateway.*

Aha	
Site search (all URLs from a particular site)	Query: related fields / subjects
See partners	*As in a directory and library gateway.*
Size, speed and scope	Relate query to best-suited search tool: database size, meta search engine, search engine, directory, specialized directory etc.
Google and ODP widen scope considerably.	*Individual assessment.*
Sorting: clustering, customization features	Stemming and truncation possibilities
Not indicated	*In addition, spelling variations.*
Special features	Synonyms
Travel Guide, News search, Aha Classifieds.	***Alternative terms.***
Spelling suggestions	**COMMENT:**
No, only partners.	*Do not use if exact search guidance and examples are required.*
Stemming	**SEARCH TOOL RECOMMENDATION:**
See partners	*Fair*
Stop words	**USER RECOMMENDATION:**
See partners	*Not for beginners needing search examples.*
Strengths	**QUERY RECOMMENDATION:**
Translation features	
See partners	
Truncation	**GENERAL RECOMMENDATION:**
See partners	*Least favorite of the three local search engines used due to its inadequate results, precision and slow response times.*
URL search (find a single page)	
See partners	
Weaknesses	
The South African Web searches only domain names ending in .za, and unlike, unlike Ananzi hat consider individual SA sites, and presents also SA sites that do not end in .za. Very slow response times. Inadequate search help for beginners.	
COMMENT:	
The use of Google and ODP increases the credibility of results, the filters might be over sensitive and filter out required information on sensitive topics.	

17.2.2 Aardvark

Web address: www.aardvark.co.za

Table 12: Search tool evaluation: Aardvark

Aardvark	
SEARCH TOOL	**USER**
Inconsistencies	Filters available
Use of search operators did not change the search results or relevancy.	*Not indicated*
Languages available	Special user requirements
Babel Fish: AltaVista translation service.	
Documentation: copyright information, Spam policy, privacy policy and protection	Search user friendliness and ease of use
Copyright information.	*User friendly, search help is excellent.*
Classification	Search planning and strategy
Directory with search engine abilities	
Default operation	Search options required: keyword / subject
Display: results, # of hits, text from hits, customization, "More links like these"…", query biased summary	Search customization features
Displays the total number of hits, 10 per page, with title, a short description, categories, the URL, cached pages and rank within results. "Search within results "option available.	*Enabled, see SE limits*
Freshness: spidering frequency	Portal / hub features offered: news, email, customization
Not indicated	*No*
Duplicate detection	Protection: privacy and Spam
No duplicates noticed during the test run.	*Not mentioned*
Field searching: URL, title, domain, etc.	Search speed
African domains	*Reported to be the fasted search engine in Africa.*
Family filters	Pop-up ads
Not indicated	*None, banner ads on right hand side of the page*
Case sensitivity	Knowledge of field of search
No	
Boolean searching	Browser window layout, user friendliness, interface
"all the words", "exact phrase", "at least one of", "without", + forces search on stop words.	*Browser window is more cluttered than Aha's home page, and the search button is less prominent. The directory is displayed on the home*

Aardvark	
	page.
Databases and searches	General search proficiency, characteristics, cognitive behavior, sensorimotor skills
Google, African Search (search for web sites on the African continent and in the African Directory), Directory search (search African directory), WWW search (search on the WWW). *During the test run, this search retrieved a fair amount of hits, and all were (within the test run parameters) relevant. .*	
Cached pages	Internet search experience
Enabled	
File types	Speed of downloads – graphics, text
Any	*Graphics and frames were cut to enhance speed.*
Limits	Relate user to best suited search tool: search engine, directory, portal, intelligent agent
Advanced search: Language, file format, date, term position in page, domain.	
Overlap,: database and source sharing	**COMMENT: easy to use, even beginners should have no problem following the easy directions for advanced searches. The Google partnership boosts confidence in the relevance of the search results.**
AltaVista translation, powered by Google.	**QUERY**
Ownership – points to possible affiliations and independence	Alternative spelling
Provided and maintained by Telkom SA LTD.	
Partner changes	Authority of the sources
Partner size comparisons	Database freshness required?
Google is the largest SE	
Phrase search	File types most likely to contain information
Enabled in advanced search	
Portal features	Filters
Favorite sites, IT news, top searches, weather, poll, etc.	
Proximity searching	Format: text, image, sound, video, presentations

Aardvark	
Quality control: moved pages, dead links, recall and precision	Information required: current and up to date
Highest recall/precision during test run, within the test run parameters.	
Ranking algorithm Paid services, keyword frequency, links frequency, "authority sites"	Information required: general and popular
Advertising banners on the right hand side of the page. Search rank displayed on results page.	
Search help	Information required: very specific and exact matches
Link is at the bottom of the home page. Search help is extensive and cater for the inexperienced and experienced searcher.	
Search options offered Basic / advanced	Overall general, broad field of query
Search and Advanced search and cached pages.	
Site search (all URLs from a particular site)	Query: related fields / subjects
Enabled in advanced search	
Size, speed and scope	Relate query to best-suited search tool: database size, meta search engine, search engine, directory, specialized directory etc.
Use of Google and AltaVista enhances reach and speed. Reported to be Africa's fastest search engine. Graphics and frames were cut to speed it up. The time it took to complete a search is indicated. *During the test run, this was the fastest search tool.*	
Sorting: clustering, customization features	Stemming and truncation possibilities
Ranked according to relevancy.	
Special features	Synonyms
Translation options	
Spelling suggestions	**COMMENT:**
Enabled	
Stemming	**SEARCH TOOL RECOMMENDATION:**
Not indicated	*Favorite search tool of the three local one tested.*
Stop words	**USER RECOMMENDATION:**

Aardvark	
Ignored, + forces the use of stop words.	*From beginner – advanced*
Strengths	**QUERY RECOMMENDATION:**
Fast search tool with apparent high recall/precision ratio with extensive help section. Cached pages. Google powered	
Translation features	
Babel Fish uses AltaVista	
Truncation	**GENERAL RECOMMENDATION:**
Not indicated	*Highly recommended.*
URL search (find a single page)	
Enabled in advanced search	
Weaknesses	
COMMENT:	

17.2.3 Ananzi

Web address: http://www.ananzi.co.za

Table 13: Search tool evaluation: Ananzi

Ananzi	
SEARCH TOOL	**USER**
Inconsistencies	Filters available
Use of search operators did not change the search result.	*None mentioned*
Languages available	Special user requirements
International Start pages available in Dutch, English (UK, Canada and USA), German and Belgium.	
Documentation: copyright information, Spam policy, privacy policy and protection	Search user friendliness and ease of use
Copyright information, FAQ, and privacy policy, webmaster and other staff contact details and names supplied.	*Easy to use, but the home page is very cluttered with banner ads.*
Classification	Search planning and strategy
Search engine with directory	*Easy to use sitemap that lists available categories and sections.*

Ananzi	
Default operation	Search options required: keyword / subject
If no field search is indicated, the text is searched for in the Title, Summary and Body. Default operator is OR.	*Different options are explained.*
Display: results, # of hits, text from hits, customization, "More links like these"…", query biased summary	Search customization features
Displays their unique hits1st, indicates the total number of matching results. Priority listings are indicated. The matching hits are numbered, ands shows the title, URL, one or two sentences of description, and the search engines where the hit (as well as the number of hits per SE) was found. "Find similar" links enabled. Sponsored links and high ranking advertised.	*Many options provided with easy to follow instructions.*
Freshness: spidering frequency	Portal / hub features offered: news, email, customization
Every 10 to 16 days.	*Enabled: email, news, etc.*
Duplicate detection	Protection: privacy and Spam
Not indicate and no duplicates found during the test run.	*Privacy protection explained.*
Field searching: URL, title, domain, etc.	Search speed
Link, site, url, title. Ananzi claims to be devoted to South African web sites, even those that have domains that do not end in .za.	*Claims to be one of the fastest Search Engines on the Internet.*
Family filters	Pop-up ads
Not mentioned	*None, but many banner ads present.*
Case sensitivity	Knowledge of field of search
Lowercase terms match any case, otherwise case is matched exactly as typed. Proper names should be capitalized.	
Boolean searching	Browser window layout, user friendliness, interface
Double quotation marks around more than one terms indicates that the words should be adjacent (Boolean ADJ), + (Boolean AND) requires a term, - (Boolean NOT) excludes documents with the term, fieldname: specifies that the term must be found in that field, | (pipe symbol) between queries, e.g.	*Search help is easy to find and follow search help, and the different indexes are well explained. Contact details are supplied. Invites comments on each search.*

Ananzi	
query1 | query 2 will search the results of query 1 with query2, ranking results by relevance to both query1 and query2. Where as in query1 | | query2 searches the results of query1 with query2, ranking results only by relevance to query2.	
Databases and searches	General search proficiency, characteristics, cognitive behavior, sensorimotor skills
Inktomi Search, SA Web, SA Site Directory, International, Ixquick Priority listings. Searches GigaBlast, AskJeeves/Teoma, FindWhat, LookSmart, Overture, Go, ODP, MSN, and Brabys. Sites are individually checked to include SA sites that do not end in .za. This was the "average" search tool in terms of results during the test run.	
Cached pages	Internet search experience
Not mentioned	
File types	Speed of downloads – graphics, text
MP3, pictures, Adobe PDF is requested.	*Fast download, no frames or graphics*
Limits	Relate user to best suited search tool: search engine, directory, portal, intelligent agent
Field searches	
Overlap: database and source sharing	**COMMENT:**
Inktomi Search	**QUERY**
Ownership – points to possible affiliations and independence	Alternative spelling
Powered by Verity and Ixquick.	*Spelling suggestions*
Partner changes	Authority of the sources
Member of the Online Publishers Association.	
Partner size comparisons	Database freshness required?
ODP is regarded as one of the biggest Internet directories.	
Phrase search	File types most likely to contain information
Double quotation marks, most accurate way to search.	
Portal features	Filters
Chat, email, news, entertainment, shop, persfin and sport.	

Ananzi	
Proximity searching	Format: text, image, sound, video, presentations
With phrase searches	
Quality control: moved pages, dead links, recall and precision	Information required: current and up to date
It is stated that Ananzi will return relevant results even if they do not contain all query terms – this would result in high recall, but low precision. During the test run Ananzi retrieved the most documents but with the least precision. This might have been different if search operators had been used to limit the search.	
Ranking algorithm Paid services, keyword frequency, links frequency, "authority sites"	Information required: general and popular
Returns results sorted with the "best" matches at the top. Inktomi weighs terms on their statistical uniqueness, and ranks them accordingly. *In the absence of other information, Ananzi Search indexes all the words in a document except comments. The1st couple of words is used as a summary.* *Paid listings appear on the right hand side of the result pages, and more are invited, presumably to appear in the ranked pages.*	
Search help	Information required: very specific and exact matches
Extensive sections on search help, explaining search syntax, meta tags, requiring and excluding terms, etc. quick tips and examples are offered.	
Search options offered Basic / advanced	Overall general, broad field of query
Search, Advanced search, Search within results (button or use the pipe symbol |). The SA Directory headings and categories are displayed on the home page.	
Site search (all URLs from a particular site)	Query: related fields / subjects
Enabled	
Size, speed and scope	Relate query to best-suited search tool: database

Ananzi	
	size, meta search engine, search engine, directory, specialized directory etc.
Claims to be one of the fastest search engines on the Internet. Currently indexes over 300 000 web pages within SA, growing daily. The Ananzi SA Site Directory is a handpicked category based list of the best sites in SA. It was one of the slowest search tools during the test run.	
Sorting: clustering, customization features	Stemming and truncation possibilities
Best matches are displayed on top.	
Special features	Synonyms
Use of pipe symbol (|) to indicate search within results and ranking accordingly. Search for alphanumeric terms and use diacritics are enabled.	
Spelling suggestions	**COMMENT:**
Not mentioned	
Stemming	**SEARCH TOOL RECOMMENDATION:**
Search results will include any variation of that word, part of normal search, not advanced search.	*Fair, must apply search operators to search syntax.*
Stop words	**USER RECOMMENDATION:**
Ignored	*Beginners to advanced.*
Strengths	**QUERY RECOMMENDATION:**
Detailed explanation of how to search this search engine, focus on SA sites, partnership with Inktomi. The pipe symbol is a unique search feature.	
Translation features	
Not indicated	
Truncation	**GENERAL RECOMMENDATION:**
As for stemming	*A fair search tool with many special features and applying the search operators should downscale the "hits overload".*
URL search (find a single page)	
Enabled	
Weaknesses	

Ananzi	
Possible skewed Recall/Precision ratio – too many hits, not precise matches. No cached pages capability. High recall but low precision ratio. The use of Ananzi operators should remedy this to some extent.	
COMMENT:	
Fair search tool with unique search features.	

17.3 Results and test run conclusion

It was interesting to compare the local search tools with the benchmark set by well-known, popular and major international search tools.

The local search tools measured up well against the major ones, and indeed all three have major international partners. In this sense then, they are no different really from other, international, smaller search engines that also collaborate with major search tools, but providing a local focus.

In terms of overlap, Aha, Ananzi and Aardvark had some unique hits that were not found in Google and Yahoo. All the Google and Yahoo hits were present in one of the local search engines, and there was much overlap between the Yahoo and Google results.

Following this test run of searches on the three local search engines it is concluded that, although they compare fairly favorably with the major search tools, one could probably find similar results using major search tools. The search on major search tools could be further limited by country.

Note that this test run was extremely limited in scope and depth, and served only as a possible indicator of trends. Further investigation is required to validate possible trends mentioned here. For the sake of this study, the conclusions reached in the next chapter will be deemed applicable to local search engines as well.

18 GENERAL CONCLUSION

Studying different major search engines, describing, and comparing their desirable features proved useful in setting basic search engine benchmarks. The benchmarks include the 'three major role players' in conducting a web based search using a search engine: the search tool, the user and the search query. These desirable features formed the basis of the search engine/tool evaluation instrument.

The world of the Internet and Web search tools is dynamic, and partners and features change almost daily, making it very challenging to keep up with new developments. The instrument for search tool evaluation developed in this study is a useful tool to assist users when selecting a new search tool because it considers not only the search tool features, but also match it with the query and the user's needs and experience. Notwithstanding developments, this instrument will be able to guide searchers as to desirable features that will enhance their search results.

The instrument developed to evaluate search tools proved useful in gaining insight into the workings of three local search tools as compared with international benchmarks. Generally, the local search tools compared favorably with the benchmark set by major search tools. Despite some unique hits by some of the local search tools, this author is not convinced that local search tools are preferable to international search tools concerning local information. Major, international search tools cover the local South African and African Web as well. Using international search tools and limiting the search by country for local information is likely to cover and include results from local search tools, but with the added benefit of the support of major search tools, development and research. Further study is required to confirm this.

In the mean time, applying the evaluation instrument when selecting search engines / tools will optimize search results and enable researchers to identify the best search engine (local or international) to obtain reliable, valid and current information faster to optimize the Recall / Precision ratio.

The development of evaluation methodologies for search engines remains a challenging but important area of research that is directly relevant to innovative design of search engines (Gwizdka & Chignell, [1998]).

This author hopes that this study will have diagnostic utility in suggesting areas for improvement of web search engine design and provide guidance on how to modify search algorithms and

presentation techniques, as well as highlight the critical points of consideration of the users' varied needs and capabilities.

19 FINAL THOUGHTS

Notess (2003n) speculates that the next year will bring about many changes to Web searching. He foresees consolidation and possibly vanishing features and databases. Consolidation will happen at the corporate level, and although most databases are still separate, some aspects have merged already (AltaVista and AlltheWeb, Yahoo already owns Inktomi and Overture). Consolidation will have a major impact on the search interface of search engines. These changes will affect searchers. Some databases could be lost, and features like truncation and proximity searching might be at risk. More documents will be accessible – more file types, Usenet at Google, Current News, books and product information. Documents might also be lost due to excessive Robots.txt exclusions and non-optimized database driven sites (many are indexed, others are not). Notess is especially concerned about the search "improvements" target that is aimed at the non-professional searcher, and not professional searchers. Search engines have diverse audiences with divergent preferences - what will be the standard?

Dalio (2003) reported on the 12th International World Wide Web Conference held in 2003 in Budapest that the search engines of the future will be fast, pretty, and personalized to suit every user's needs. Search engines have improved significantly since the beginning of the Web, and yet they look and function very much the same as ever. Computer scientists are working on new search techniques and user interfaces that could significantly alter most searchers' result pages.

Searchers might in future be able to sift through search results graphically, or personalize results. Such specialized search tools that work on one computer or across a private network are a step closer toward very personalized searching of the Internet, where results will be tailored to the express desires of each user.

Search tools will eventually be enabled to consult cookies (small text files that the Web server placed on the hard drive of the computer used by the searcher) and assume from past searches, that a searcher is looking for one type of information and not another. These cookies could tell the search engine to return only new information, or only data targeted to the user's location (Dalio, 2003).

The author suggests that Intelligent Agents will soon be the new way to search, update and organize information on the Web and Internet. Already such software can be installed on the searcher's computer. Soon enough, searchers and users alike will be able to use even the most cumbersome search tool (if such a search tool survives in this competitive field) to find even the

most elusive of information available on the Internet. It is likely that individual search engine features and specifically their user end – the browser interface and user-friendliness might become less important, whereas search engine software, their indexing ability and scope might still be reflected in the Recall/Precision/Response Time ratio.

This advancement raises another issue that is becoming increasingly important – the protection of the searcher's privacy. A cookie that stores information on past searches and preferences and that is enabled to make this information available to a search engine requesting it, might not necessarily be in the best interest of the searcher. Indeed this opens a new field of research, developments and innovation.

20 SOURCES CONSULTED

Aardvark. c2004a. Aardvark advanced search [online] Available: http://www.aardvark.co.za/search/AdvancedSearch.php [9 January 2004]

Aardvark. c2004b. Aardvark help [online] Available: http://www.aardvark.co.za/search/help.php [9 January 2004]

Aardvark. c2004c. About us [online] Available: http://www.aardvark.co.za/search/about_us.php [9 January 2004]

Aardvark. c2004d. Ask an aardvark [online] Available: http://www.aardvark.co.za/search/ [9 January 2004]

Aha. c2003a. About us and terms of use [online]. Available: http://www.aha.co.za/webcentre/about_us.asp [11 January 2004]

Aha. c2003b. RSA Directory / Top [online]. Available: http://www.aha.co.za/directory/top.asp [11 January 2004]

Aha. c2003c. Site map [online]. Available: http://www.aha.co.za/webcentre/site_map.asp [11 January 2004]

Aha. c2004a. Aha! – RSA Directory / Search [online]. Available: http://www.aha…./sawsearch.asp?keywords=south%20africa%20hippotherapy&submit=Search&h= [11 January 2004]

Aha. c2004b. Search tips [online]. Available: http://www.aha.co.za/webcentre/searchtips.asp [11 January 2004]

Aha. c2004c. The Great South African Internet Directory [online]. Available: http://www.aha.co.za/default.asp [11 January 2004]

AltaVista. c2004. AltaVista – basic help [online]. Available: http://www.altavista.com/help/search/default [6 January 2004]

Ananzi. c2003c. Meta tags [online]. Available: http://search.ananzi.co.za/help/meta.html?la=en [12 January 2004]

Ananzi. c2003d. Quick tip and examples [online]. Available: http://search.ananzi.co.za/help/la=en [12 January 2004]

Ananzi. c2003e. Requiring or excluding terms [online]. Available: http://search.ananzi.co.za/help/boolean.html?la=en [12 January 2004]

Ananzi. c2003f. Search syntax summary [online]. Available: http://search.ananzi.co.za/help/syntax.html?la=en [12 January 2004]

Ananzi. c2003g. Special searches [online]. Available: http://search.ananzi.co.za/help/special.html?la=en [12 January 2004]

Ananzi. c2003h. Startpage [online]. Available: http://www.startpage.co.za [12 January 2004]

Ananzi. c2004a. About Ananzi [online]. Available: http:search1.ananzi.co.za/About_us/ [12 January 2004]

Ananzi. c2004b. Ananzi South Africa – search engine [online]. Available: http://www.ananzi.co.za [12 January 2004]

Ananzi. c2004c. Ananzi tips [online]. Available: http://www.ananzi.co.za/today/tips.html [12 January 2004]

Ananzi. c2004d. How does Ananzi work? [online]. Available: http://search1.ananzi.co.za/faq/works.html [12 January 2004]

Ananzi. c2004e. Sitemap [online]. Available: http://www.ananzi.co.za/sitemap/index.html [12 January 2004]

Ayache, Michelle. 2003. Search engine marketing [online]. Available: http://www.digitalmeesh.com/sepres/SearchEngineMarketing.doc [15 December 2003]

Barlow, Linda. 2002a. A helpful guide to web search engines [online]. Available: http://www.monash.com/spidap4.html [21 December 2003]

Barlow, Linda. 2002b. Frequently asked questions about search engines [online]. Available: http://www.monash.com/spidap2.html [21 December 2003]

Barlow, Linda. 2002c. How to plan the best search strategy [online]. Available: http://www.monash.com/spidap1.html [21 December 2003]

Barlow, Linda. 2002d. The web search wizard [online]. Available: http://www.monash.com/spidap5.html [21 December 2003]

Bharat, Krishna & Broder, Andrei. 1998. A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems,* 30(1998): 379-388.

Carr, Maureen, Santowski, Britt & Marzolf, Jennifer. 2000. Search engine types [online]. Available: http//216.239.37.104/se.../ep0504g_v01.htm+%2B%22search+engine+types%22&hl=en%ie=UTF- [14 December 2003]

Chu, Heting & Rosenthal, Marilyn. 1996. Search engines for the World Wide Web: a comparative study and evaluation methodology [online]. In ASIS 1996 Annual Conference Proceedings October 19 – 24, 1996. Available: http://www.asis.org/annual-96/ElectronicProceedings/chu.html [11 November 2003]

Collier, Harry & Arnold, Stephen E. 2003. Search engines: evolution and diffusion. Harrod's Creek, Kentucky.

Common resolutions. 2002. By Katherine A. Stalcup...[et al.] [online]. Available: http://www4.tltc.ttu.edu/search/common.htm [14 December 2003]

CompInfo – The Computer Information Center. c2002. Intelligent agents [online]. Available: http://www.compinfo-center.com/tpaqnt-t.htm [29 December 2003]

comScore Media Metrix Search Engine Ratings. 2003. Edited by Danny Sullivan. [online] Available: http://www.searchenginewatch.com/reports/print.php/34701_2156431 [21 December 2003]

Corman, Sheila. 2002. Book review: The extreme searchers guide to web search engines: a handbook for the serious searcher by Randolph Hock, 2001. *Serials Review,* 28(1): 83-84.

Cricket Software Limited. c2003. Fact sheet: Listed on search engines [online]. Available: www.crickett.co.uk [14 December 2003]

Dalio, Michelle. 2003. Big changes for search engines. Wired news, 27 May [online]. Available: http://wired.com/news/technology/0,1282,58971,00.html [4 January 2004]

Day, James. [2001]. The quest for information: a guide to searching the Internet, 2(4) [online]. Available: http://www.thejcdp.com/issue008/day/04day.htm - http://www.thejcdp.com/issue008/day/11day.htm [14 December 2003]

Ding, Ying, Chowdhury, Gobinda & Foo, Schubert. 2000. Organizing keywords in a Web search environment: a methodology based on co-word analysis. In *Dynamism and stability in knowledge organization – Proceedings of the Sixth International ISKO* Conference (2000: Toronto). Toronto: Ergon Verlag. p. 28-34.

FAQ's. 2002. By Katherine A. Stalcup...[et al.] [online]. Available: http://www4.tltc.ttu.edu/search/FAQs.htm [14 December 2003]

Ferguson, Janet. 2003. Evaluating World Wide Web information [online]. (Ramsey Library Research Guides) Available: http://bullpup.lib.unca.edu/library/lr/evalweb.html [14 December 2003]

Google. c2003a. Advanced Google search operators [online]. Available: http://www.google.com/help/operators.html [7 December 2003]

Google. c2003b. Advanced search made easy [online]. Available: http://www.googl.com/help/refinesearch.html [21 December 2003]

Google. c2003c. Customize your Google search results [online]. Available: http://www.googl.com/help/customize.html [21 December 2003]

Google. c2003d. Google help [online]. Available: http://www.google.com/help/basics.html [21 December 2003]

Google. c2003e. Google help central [online]. Available: http://www.google.com/help/index.html [21 December 2003]

Google. c2003f. Google services and tools [online]. Available: http://www.google.com/options/index.html [21 December 2003]

Google. c2003g. Google technology [online]. Available: http://www.google.com/technology/index.html [21 December 2003]

Google. 2003h. Google web search features [online]. Available: http://www.google.com/help/features.html [21 December 2003]

Google. c2003i. How to interpret your search results [online]. Available: http://www.google.com/help/interpret.html [21 December 2003]

Gray, Terry A. [date unknown]. How to search the web: a guide to search tools [online]. Available: http://daphne.palpmar.edu/tgsearcg/ [8 December 2003]

Gwizdka, Jacek & Chignell, Mark. [1998]. Towards information retrieval for evaluation of Web search engines. Toronto: University of Toronto.

Habib, David & Balliot, Robert L. 2003. How to search the World Wide Web: a tutorial for beginners and non-experts [online]. Available: http://middletownpubliclibrary.org/tutor.htm [18 December 2003]

Hawking, David, Craswell, Nick & Griffiths, Kathleen. [2000]. Which search engine is best at finding online services? Canberra: CSIRA Mathematical and Information Sciences & Centre for Mental Health Research, Australian National University.

Haynal, Russ. c1999. An overview of search tools [online]. Available: http://navigators.com.search.html [21 December 2003]

Hölscher, Christoph & Strube, Gerhard. 2000. Web search behavior of Internet experts and newbies. *Computer Networks,* 33(2000): 337-346.

Indiana University. 2003. Evaluating web sites [online]. Available: http://www.indiana.edu/~libugls/Publications/webeval.html [14 December 2003]

Intelligent Agents. [date unknown]. Available: http://www.intelligent-agents/com [29 December 2003]

Internet web draughting. 2002a. Search engine introduction [online]. Available: http://iwd.co.za/search_engines/searchengine_introduction.htm [14 December 2003]

Internet web draughting. 2002b. Search engine types [online]. Available: http://iwd.co.za/search_engines/searchengine_types.htm [14 December 2003]

IWD see Internet web draughting.

James Madison University Libraries. [2003]. Internet search [online]. Available: http://www.lib.jmu.edu/internet/ [14 December 2003]

Jansen, James. c1996. Using an intelligent agent to enhance search engine performance [online]. Available: http://www.firstmonday.dk/issues/issue2_3/jansen/ [21 December 2003]

JMU Libraries see James Madison University Libraries.

Kansas City Public Library. c2002a. AltaVista [online]. Available: http://www.kclibrary.org/resources/search/altavista.cfm [7 December 2003]

Kansas City Public Library. c2002b. Feature comparison chart [online]. Available: http://www.kclibrary.org/resources/search/chart.cfm [7 December 2003]

Kansas City Public Library. c2002c. Introduction to search engines [online]. Available: http://www.kclibrary.org/resources/search/intro.cfm [7 December 2003]

KCPL see Kansas City Public Library

Kokkelink, Stefan & Schwänzl, Roland. 2002. Expressing qualified Dublin Core in RDF / XML [online]. Available: http://dublincore.org/documents/202/04/14/dcq-rdf-xml/index.shtml [12 January 2004]

Laudon, Kenneth C & Traver, Carol Guericio. c2002. *E-commerce: business, technology, society.* [Boston: Addison Wesley].

Le Roux, André. 2003a. Search engine relationships [online]. Available: http://www.searchengineyearbook.com/search-engine-relationships.shtml [30 December 2003]

Le Roux, André. 2003b. Search engine statistics [online]. Available: http://www.searchengineyearbook.com/search-engines-statistics.shtml [30 December 2003]

Level Ten. c2002a. Internal search engine – principles, tips, example [online]. Available: http://www.leveltendesign.com/kb_terms/internal_search_engine.html [14 December 2003]

Level Ten. c2002b. External search engine – principles, tips [online]. Available: http://www.leveltendesign.com/kb_terms/external_search_engine.html [14 December 2003]

Level Ten. c2002c. Search engine - principles [online]. Available: http://www.leveltendesign.com/kb_terms_search_engine.html [14 December 2003]

Level Ten. c2002d. Search engine spider – definition [online]. Available: http://www.leveltendesign.com/kb_terms/search_engine_spider.html [14 December 2003]

Level Ten. c2002e. Web page search engine – principles [online]. Available: http://www.leveltendesign.com/kb_terms/web_page_search_engine.html [14 December 2003]

Level Ten. c2002f. Web site search directory – principles [online]. Available: http://www.leveltendesign.com/kb_terms/web_site_search_directory.html [14 December 2003]

Ljosland, Mildrid. 1999. Evaluation of Web search engines and the search for better ranking algorithms [online]. A paper given at the SIGIR99 Workshop on Evaluation of Web Retrieval, 19 August 1999. Available: http://www.aitel.hist.no/~mildred/dring/paper/SIGIR.html [11 November 2003]

Lycos. [2004]. Lycos home page [online] Available: http://www.lycos.com [6 January 2004]

Methodologies and website development: a survey of practice. 2002. By M.J. Taylor ...[et al.] *Information and Software Technology,* 44(2002): 381-391.

Notess, Greg R. [1999]. Comparing Internet search engines [online]. Available: http://www.csu.edu.au/special/online99/proceedings99/103a.htm [8 December 2003]

Notess, Greg R. [date unknown]. Definitions [online]. Available: http://www.searchengineshowdown.com/defs [8 December 2002]

Notess, Greg R. 2002a. Multiple search engines [online]. Available: http://searchengineshowdown.com/multi [8 December 2003]

Notess, Greg R. c2002b. Review of Excite [online]. Available: http://www.searchengineshowdown.com/features/excite/index.shtml [4January 2004]

Notess, Greg R. 2002c. Top tips for searchers [online]. Available: http://notess.com/speak/talks/2002toptips.pps [8 December 2003]

Notess, Greg R. c2003a. AltaVista inconsistencies [online]. Available: http://www.searchengineshowdown.com/features/av/inconsistent.shtml [8 December 2003]

Notess, Greg R. 2003b. Beyond search engines [online]. Available: http://searchengineshowdown.com/strat/advanced/searchtoc.html [8 December 2003]

Notess, Greg R. 2003c. GigaBlast adds spelling suggestions [online]. Available: http://www.searchengineshowdown.com/newsarchive/000746.shtml [8 December 2003]

Notess, Greg R. 2003d Google starts auto stemming searches [online]. Available: http://searchengineshowdown.com/newsarchive/000742.shtml [8 December 2003]

Notess, Greg R. 2003e. Review of AltaVista [online]. Available: http://www.searchengineshowdown.com/features/av/index.shtml [29 November 2003]

Notess, Greg R. 2003f. Review of Google [online]. Available: http://www.searchengineshowdown.com/features/google/index.shtml [8 December 2003]

Notess, Greg R. c2003g. Review of Lycos [online]. Available: http://www.searchengineshowdown.com/features/lycos/index.shtml [8 December 2003]

Notess, Greg R. 2003h. Review of Yahoo! Directory [online]. Available: http://www.searchengineshowdown.com/dir/yahoo/index.shtml [4 January 2004]

Notess, Greg R. 2003i. Search engine showdown reviews [online]. Available: http://www.searchengineshowdown.com/reviews/ [29 November 2003]

Notess, Greg R. 2003j. Search engine statistics: freshness showdown [online]. Available: http://searchengineshowdown.com/stats/freshness.shtml [8 December 2003]

Notess, Greg R. 2003k. Search engine statistics: relative size showdown [online]. Available: http://www.searchengineshowdown.com/stats/size.shtml [8 December 2003]

Notess, Greg R. 2003l. Search engines by search features [online]. Available: http://searchengineshowdown.com/features/byfeature.shtml [8 December 2003]

Notess, Greg R. 2003m. Web search secrets: advanced features and failures [online]. Available: http://notess.com/speak/talks/infotoday2003secrets.pps [8 December 2002]

Notess, Greg R. 2003n. Web searching in 2004 [online]. Available: http://notess.com/speak/talks/il03webwearching2004.pps [8 December 2002]

Notess, Greg R. 2003o. Why search engine size matters [online]. Available: http://searchengineshowdown.com/stats/sizematters.shtml [8 December 2003]

Pay per click search engine Internet marketing. c2002. What is pay per click advertising? [online]. Available: http://www.pay-per-click-search-engine-internet-marketing.com/what.html [14 December 2003]

Peterson, Richard Einer. c1997. Eight Internet search engines compared [online]. Available: http://www.firstmonday.dk/issues/issue2_2/peterson/ [21 December 2003]

PPC see Pay per click search engine Internet marketing.

Rogers, Ian. 2002. The Google pagerank algorithm and how it works [online] Available: http://www.iprcom.com/papers/pagerank [4 January 2004]

Results and challenges in Web search evaluation. 1999. By Hawking, David ...[et al.] *Computer Networks,* 31(1999): 1321-1330.

Savoy, Jacques & Picard, Justin. 2001. Retrieval effectiveness on the web. *Information Processing and Management,* 37(2001): 543-569.

Search engine marketing. [2002] Available: http://www.digitalmeesh.com/sepres/se-complete.pdf [21 December 2002]

Search engine marketing. [2003a]. Relationships between the major search engines [online]. Available: http://www.digitalmeesh.com/sepres/relationships.htm [14 December 2003]

Search engine marketing. [2003b]. What you need to know about search engines [online]. Available: http://www.digitalmeesh.com/sepres/terms.htm [14 December 2003]

Search engine types. 2002. By Katherine A. Stalcup...[et al.] [online]. Available: http://www4.tltc.ttu.edu/search/search_engine_types.htm [14 December 2003]

Spannerworks. 2003a. Are some search engines more important than others? [online]. Available: http://www.spannerworks-positioning.co.uk/faq/relative_importance.html [21 December 2003]

Spannerworks. 2003b. Glossary [online]. Available: http://www.spannerworks-positioning.co.uk/glossary.html [21 December 2003]

Teacher Resource Centre. c2002a. Filtered search engines [online]. Available: http://www.chinooksedge.ab.ca/teacher_centre/search/filtered.html [14 December 2003]

Teacher Resource Centre. c2002b. Hierarchical search engines [online]. Available: http://www.chinooksedge.ab.ca/teacher_centre/search/hierarchical.html [14 December 2003]

Teacher Resource Centre. c2002c. Media search engines [online]. Available: http://www.chinooksedge.ab.ca/teacher_centre/search/media.html [14 December 2003]

Teacher Resource Centre. c2002d. Natural language search engines [online]. Available: http://www.chinooksedge.ab.ca/teacher_centre/search/natural.html [14 December 2003]

Testing Labs. 2000. Google web search engine. Morrisville, NC.

The Excite Network. c2004a. Excite – search help [online]. Available: http://www.infospace.com/info.excite/about/corporate/help.htm [4 January 2004]

The Excite Network. c2004b. My Excite [online]. Available: http://www.excite.com [4 January 2004]

The web pros group. c2004. Types of search engines and opportunities [online]. Available: http://www.webprosgroup.com/search-engine-types.htm [14 December 2003]

Top 10 complaints. 2002. By Katherine A. Stalcup...[et al.] [online]. Available: http://www4.tltc.ttu.edu/search/10complaints.htm [14 December 2003]

Top Ranked Position. 2003a. How to check your URL directly on the search engines [online]. Available: http://searcg-engine-positioning-seo.com//seo/check_url.html [15 December 2003]

Top Ranked Position. 2003b. Pay per click placement [online]. Available: http://searcg-engine-positioning-seo.com//seo/resources/pay_per_click_tips.html [15 December 2003]

Top Ranked Position. 2003c. Pay per click search engines [online]. Available: http://searcg-engine-positioning-seo.com//seo/resources/pay_per_click_engines.html [15 December 2003]

Top Ranked Position. 2003d. Search engines, two basic types [online]. Available: http://searcg-engine-positioning-seo.com/search_engines.html [15 December 2003]

Top Ranked Position. 2003e. The major search engines and directories [online]. Available: http://searcg-engine-positioning-seo.com/seo/resources/search_engines1.html [15 December 2003]

TRC see Teacher Resource Centre

Wall, Aaron. [2003a]. Future of search engines [online]. Available: http://www.search-marketing.info/future-of-search-engines/index.htm [21 December 2003]

Wall, Aaron. [2003b]. Major search engines [online]. Available: http://www.search-marketing.info/search-engines/major-search-engines/major-engines.htm [21 December 2003]

Wall, Aaron. [2003c]. Other search engines [online]. Available: http://www.search-marketing.info/search-engines/other-search-engines/index.htm [21 December 2003]

Wall, Aaron. [2003d]. Search engine algorithms [online]. Available: http://www.search-marketing.info/search-algorithm/ [21 December 2003]

Wall, Aaron. [2003e]. Search software and user interface [online]. Available: http://www.search-marketing.info/search-engines-work/search-software-interface.htm [21 December 2003]

Wang, Peiling, Hawk, William B & Tenopir, Carol. 2000. User's interaction with World Wide Web resources: an exploratory study using a holistic approach. *Information Processing and Management,* 36(2000): 229-251.

Web Site Promotion Optimization Services. [2000]. The following search engines represent over 90% of all English speaking results [online]. Available: http://www.website-promotion-optimization-services.com/images/search-engine-traffic-chart.gif [21 December 2003]

WebWorkshop. c2003. Google's pagerank calculator [online]. Available: http://webworkshop.net/pagerank_calculator.php3 [4 January 2004]

Weiss, Scott. 1997. Glossary for Information Retrieval [online]. Available: http://www.cs.jhu.edu/~weiss/glossary.html [29 November 2003]

What is a search engine? 2002. By Katherine A. Stalcup...[et al.] [online]. Available: http://www4.tltc.ttu.edu/search/whatisa.htm [14 December 2003]

White, Ryen W, Jose Joemon M & Ruthven, Ian. 2003. A task-oriented study on the influencing effects of query-biased summarization in web searching. *Information Processing and Management*, 39(2003): 707-733.

WSPOS see Web Site Promotion Optimization Services

Yahoo! c2004. Yahoo! Help [online]. Available: http://help.yahoo.com/help/us/ysearch/basics/basics-01.html [6 January 2004]

Yahoo! c2004a. Search shortcuts and preferences help [online]. Available: http://help.yahoo..com/help/us/ysearch/tips/tips-03.html [6 January 2004]

Yahoo! c2004b. Yahoo! [online]. Available: http://www.yahoo.com [6 January 2004]

Yahoo! c2004c. Yahoo! Search Help [online]. Available: http://help.yahoo..com/help/us/ysearch/basics-11.html [6 January 2004]

Zaino, Jeff. [2003a]. Boolean Logic [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld020tm [14 December 2003]

Zaino, Jeff. [2003b]. Implied Boolean Operators [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld024.htm [14 December 2003]

Zaino, Jeff. [2003c]. Library Gateways [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld013.htm [14 December 2003]

Zaino, Jeff. [2003d]. Search engine ranking [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld005.htm [14 December 2003]

Zaino, Jeff. [2003e]. Search engines: Pros and cons [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld004.htm [14 December 2003]

Zaino, Jeff. [2003f]. Some well known meta-search engines [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld008.htm [14 December 2003]

Zaino, Jeff. [2003g]. Some well-known Library Gateways [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld016.htm [14 December 2003]

Zaino, Jeff. [2003h]. Some well-known subject directories [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld012.htm [14 December 2003]

Zaino, Jeff. [2003i]. Subject directories: pros and cons [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld010.htm [14 December 2003]

Zaino, Jeff. [2003j]. Troubleshooting [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld043.htm [14 December 2003]

Zaino, Jeff. [2003k]. When to use Library Gateways [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld015.htm [14 December 2003]

Zaino, Jeff. [2003l]. When to use search engines [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld006.htm [14 December 2003]

Zaino, Jeff. [2003m]. When to use subject directories [online]. Available: http://newcanaanlibrary.org/search%20engines%20with%20altavista/tsld011.htm [14 December 2003]