McMASTER
UNIVERSITY

MICHAEL G. DeGROOTE
SCHOOL OF BUSINESS

RESEARCH AND WORKING PAPER SERIES

CHARACTERIZING WORLD WIDE WEB SEARCH STRATEGIES

By

N.P. Archer

Michael G. DeGroote School of Business
McMaster University,
Hamilton, Ontario,
Canada L8S 4M4

Working Paper # 415

April, 1996

CHARACTERIZING WORLD WIDE WEB SEARCH STRATEGIES

By

N.P. Archer

Michael G. DeGroote School of Business
McMaster University,
Hamilton, Ontario,
Canada L8S 4M4

Working Paper # 415

April, 1996

Characterizing World Wide Web Search Strategies[1]

N.P. Archer
School of Business
McMaster University

Abstract

The World Wide Web (or just "Web") is a hypermedia system operating on the Internet which provides access to Web pages that exist on thousands of servers throughout the world. The growth of the Web has been phenomenal, with approximately 100 million pages currently available. To find relevant information in this huge distributed database without being overwhelmed by information overload requires strategies which make appropriate use of the logical links among these Web pages, along with the directories and search engines which have been developed to support Web search. We describe five search objectives: exploration, known-item search, topic search, general search, and exhaustive search. Web search strategies are then described for finding the most relevant information in each situation and thus meet these objectives in the most effective manner. A model is described which can be used to estimate the total number of relevant Web pages on a particular topic, given the search results and overlaps in retrievals from two or more search engines on that topic. This model is particularly useful in exhaustive search mode, and an example of its use is demonstrated with a real Web search that uses the results from five search engines.

---

[1]This work was supported through a grant from the Social Science and Humanities Research Council of Canada.

Table of Contents

1. Introduction

A large fraction of the new information now being created in business, industry, education, the arts, and recreation, is in digital form (computer created documents, digital audio recordings, digital video recorders, photoCD, digital scanners, etc.) and the rate of retrospective conversion from other forms to digital is large and growing. The attractiveness of digital encoding is that digital information can be represented as electronic, optical, and magnetic phenomena and thus stored, processed, manipulated, and transmitted readily. The concurrent and rapid growth in the availability of both public and private digital communication networks and their global interlinking (through Internet, a "network of networks"), has created a significant change to the environment in which business, personal, and professional activities are conducted. Not only have new forms of information recording and representation been created, but the speed at which this information can be transmitted to end-users has greatly reduced delays in accessing information. In the business world especially, this has shortened the time in which processes can be carried out or decisions made, with the potential for creating competitive advantage. On the other hand, improved access, faster speed, and lower cost is working to level the playing field among companies of all sizes, in reaching the global marketplace.

Direct communication to support immediate needs can be handled easily over electronic networks through forms such as electronic mail for person to person contact, or related group support systems such as list servers which broadcast communications to groups of individuals. However, the bulk of electronic text information is located in digital libraries on servers where it can be accessed on demand from client sites. One type of such libraries is represented by the information abstraction services available online, in such systems as Medline, ABI/Inform, Medlars, etc. (Pao 1989). These systems typically provide abstracts and titles of published books, articles, and other references. Most have been indexed manually and made available through remote communication links or in some cases through CDROM (e.g. ABI/INFORM for business-oriented journal titles and abstracts). Similar forms exist for computer-based library catalog systems. But there has been a recent explosion in other forms of document-based information

that are available electronically through Internet access. The most rapidly growing system that supports such information access to a variety of media forms is the World Wide Web (or just plain Web), a hypermedia system that provides access to "pages" stored on Web servers for internal organizational ("Intranet") or external (Internet) user access. As of this date (Spring 1996), there are approximately 100 million Web pages (estimates vary widely; nobody knows for sure) on many thousands of Web computer servers connected to the Internet and distributed around the world, available to millions of Internet users (e.g. some 40 million people in North America). In effect, a vast distributed database is being created, with a minimum of centralized standardization and management.

The type of information available on the Web includes information for personal interest or entertainment, and business, professional, and educational applications. The extremely rapid growth in the availability of information through the Web is an indication of the contagion stage of this system's usage, but it also indicates that we are at a point where the amount of information available is outrunning our capability to make appropriate use of it. With so much information available, it is of little use unless it can be catalogued and searched effectively in a manner that allows prospective users to retrieve relevant information easily. To assist in finding relevant information on the Web, there are a number of search engines available that can be accessed via their Web pages, along with a large number of directories that contain helpful information on links to specific topic areas. As an indication of the magnitude of Web use, Alta Vista is a Web search engine which is claimed to index over 20 million Web pages in its database, and is accessed over 5 million times each working day. Clearly, there are a very large number of users who make use of search facilities to locate Web information.

Until now, little has been done to characterize Web search strategies except for a study by Catledge and Pitkow (1996), which captured user events and related them to navigation strategies. The Web supports browsing strategies that are attractive to end-users but may be inefficient for fact retrieval. Research related to on-line text searching (Pao 1989; Blair 1990) usually focuses on systems that assist intermediaries in finding documents that satisfy criteria (key

words, associations, etc.), which have been specified by end-users. Intermediaries are experienced users of search techniques who can assist end-users to find relevant and timely, information, and they are more likely than end-users to focus on a specific goal, in order to yield efficient and cost-effective performance. They are also likely to retrieve more of the relevant information (along with possibly irrelevant material) because of their experience and ability in focusing on the task. This directly impacts the efficiency of the search. On the other hand, end-users are more likely to browse and perhaps pick up useful bits of information unrelated to the original objective of the search. In searching for a specific item, the user will know immediately when the appropriate information is found, whereas the intermediary may not. For an exhaustive topic search, unless a properly subject-indexed collection is being searched, there is no way for any user to conclude when a search is complete.

Most development of automated search support and indexing (as provided through search engines on the Web) is on systems that can support end-users and replace intermediaries in providing an analytical focus to the task, thus retrieving documents or records that satisfy some pre-determined criteria. This is the antithesis of browsing, where end-users may use inefficient information-seeking strategies in exploratory ways in hopes that information found incidentally will be useful to them; but this is typical of ill-defined problems and for learning about new task domains (Marchionini & Shneiderman 1988). The Web user may choose either to access search engines and analytical search support or to browse endlessly. But in order to make effective use of the huge Web data base and at the same time to avoid serious information overload, it is necessary to develop strategies for searching that will either bring relevant information forward very quickly if it exists, or give an indication of its non-existence if it is not available. Otherwise, the searcher can waste a great deal of time looking for information which may be available but is not being found due to a lack of knowledge about how to make the appropriate enquiries.

The objective of this study is to discuss the current state of Web information and the tools available to end users to find and retrieve that information, in the light of a variety of personal, professional, industrial, and business information needs, and to suggest appropriate strategies for

finding and retrieving this information. The question of access, cost, and payment for these services is an important issue that will need to be addressed as the information available on the Web continues to grow in volume and value to users. However, this question will not be addressed in this study.

The remainder of this document first introduces the nature of information and its uses, and explains the World Wide Web and its associated information sources. Information search strategies are then explored, and various approaches to information search on the Web through page references, directories, and search engines are described. Search effectiveness measures for information retrieval are explained, and a detailed example of exhaustive search is developed which applies these and other new measures relating to search engine performance. Finally, a series of recommended Web search strategies is outlined, followed by a discussion of the future and utility of Web information.

2. The Nature and Uses of Information

Information is data that is relevant in a particular situation and is presented in a form that is meaningful to the recipient (Senn 1990). There are a number of attributes which can be used to describe information, and these determine its usefulness. In particular, an *Item of Information* has attributes: *Scope*, *Accuracy*, *Form*, *Origin*, *Time Horizon*, and *Frequency*. Some of these attributes are immediately apparent to the recipient since they can usually be determined from the information item itself (Scope, Form, Time Horizon, and sometimes Origin). But accuracy in particular is difficult to determine directly, and gives rise to uncertainty where important decisions rely on its accuracy. The attributes of a *Set of Information*, which is primarily the focus of this discussion, are: *Relevance*, *Completeness*, and *Timeliness*. Relevance is the key factor in differentiating information from data. In the Internet world, although there is a vast amount of information available, it is almost all irrelevant to any particular user who has a certain topic in mind at any particular time. This is not unlike going into a public library, where users need support in focusing on the location(s) where relevant information is likely to be found.

On the Web, unassisted end users can be totally overloaded with data, and the key questions are: a) to find the tools and techniques that will help to differentiate information from data, and b) to provide a set of information which is complete and timely enough to be satisfactory to the user. There are some situations where it is necessary to retrieve all the relevant information available (exhaustive search), but this is not always necessary nor desirable. Users are often willing to compromise or "satisfice" (Simon 1960) with less than the complete set of information. That is, they stop searching for information when the amount of relevant information they have is satisfactory but not necessarily complete. Timeliness is often of concern to Web users, who will be more likely to access "primary" Web information (information presented directly on Web pages) if they believe this is the latest information available. Retrieved Web information is not usually wanted as an end in itself but as a necessary means to solve a problem or to continue an activity which has been arrested because the necessary information is lacking. In this case only enough relevant and timely information is needed to solve the problem at hand or to continue in the current activity.

Generally, at least five tasks can be identified (Sundaram 1995) that give rise to a need for information search. These include: problem solving, decision making (environmental scanning, developing alternative solutions, evaluating alternatives, choosing the "best solution") (Simon 1960), learning (initial learning, expanding/extending, focusing, clarifying, reviewing), calculation, and verifying. These tasks may be involved in any particular search sequence, and users may switch from one type of task to another during a session. The following sections describe the World Wide Web and then expand on how information search can be applied most effectively on the Web for a variety of such tasks.

3. The World Wide Web

The World Wide Web (Berners-Lee et al 1994) is a hypermedia system that operates over the (public) Internet, through commercial on-line service providers, private value-added networks such as America Online, Compuserve, Prodigy, etc., and through gateways to internal company

networks ("Intranets") where information providers make Web pages available. Web pages are normally in hypertext markup language (HTML) form (Aronson 1994), which includes text and references to images, sound, video, and other media forms, on host computers that may be inter-connected to networks throughout the world. Web pages are relatively simple to create through HTML (HyperText Markup Language) which is a subset of SGML (Standardized General Markup Language), and these pages are a cost effective way of making current information generally available for network access, either on public or private networks. The Internet addresses of these pages are referred to as URLs (Uniform Resource Locators)[2]. Users retrieve copies of these pages and their associated image, voice, etc. files, for viewing on their own (client) machines by using Graphical User Interface (GUI) browsers, communicating over networks to host sites through HyperText Transfer Protocol (HTTP) (Berners-Lee et al 1994). Pages referenced elsewhere, on the page being viewed, may also be retrieved for viewing, simply by clicking on a highlighted text, icon, or image, thus generating an access request to that page's URL, resulting in its retrieval. There are a variety of Web browsers available, including Mosaic, Spyglass, Netscape, Win-Tapestry, and Web-Explorer (Berghel 1996). Lynx is a widely used line-mode browser which retrieves only text and consequently requires much less communications bandwidth than graphical browsers.

The motivation for providing information as hypermedia through the Web arose initially in the academic world, but in the past few years there has been a wide-spread adoption of Web applications in business. The analogy one can draw between Web applications and other forms of business communication is that the Web is a "pull" form, where information is passive and not retrieved from the host machine unless the user wants to see it, very much like magazines, journals, and catalogs. This contrasts with television and radio for example, which are "push" forms that provide information continuously to the user. Included in these forms is advertising, superimposed on other information the user wishes to see. Much of the business information and

---

[2]URLs appear (in angle brackets) throughout this document. They are identifiable by the first few characters, which are "http://" for Web pages. Gopher sites (text only) which can be reached by Web browsers are identified by "gopher://" .

services which can be retrieved "free" on the Web are usually accompanied by advertising which helps providers to recover the associated costs. It should be noted that the passive nature of Web information (including associated advertising) has made this medium acceptable to the academic and professional community who until recently were the only users of the Internet. They do not wish to be bombarded with a continuous stream of advertising, but seem to be willing to accept it when it is directly related to information being retrieved from the Web.

The growth of Web activity is well-documented (e.g. Berghel 1996), and it now accounts for more of the data packets moved over the Internet than any other system, and the number of Web pages devoted to commercial applications is currently doubling about every two months.

3.1 Information Sources on the World Wide Web

Generally, the information directly available on Web pages includes:

- Personal (Health, Entertainment, Education/Training, Product & Service Acquisition)
- Business (Environmental Scanning, Education/Training, Communication, Product & Service Acquisition, Sales, Service and Advertising)
- Professional (Environmental Scanning, Education/Training, Communication, Publication, Product & Service Acquisition, Advertising, Sales & Service)

Such information will be referred to as *primary information*. It covers a broad range of topics, and can be very useful in support of learning, entertainment, decision making, and business transactions. An example of a Web page directory which links to a large number of current primary business information sources is <http://www.dks.com/dks/businessworld/news.html> (includes links to Associated Press, CBS Sports, CNN World News, CNN Weather, The Economist, Byte Magazine, etc., etc.).

Secondary information is information contained in databases which can also be accessed via Web pages, but the contents of these sources are usually not indexed by Web search engines (described in a following section). Through these databases one can find reference material and

factual information on history, geography, science, engineering, business, etc., in a broad range of topics which have been researched and published elsewhere, but made available by links to the Web. Much of this information has been converted retrospectively through digitization or made available from existing databases. This provides links to some of the information which predates the Web, and also provides access to some current digitized information created outside Web pages. Direct access to secondary information in most journals, books, etc. which a) has accumulated over time, b) comprises the vast majority of knowledge in the world today, and c) is also important in such activities as learning and decision making, will continue to improve via the Web. Examples of a few secondary Web sources include:

- Brittanica Online <http://www.eb.com/eb.html> (an excellent reference which includes a natural language interface, but there is a charge for use beyond demonstration retrievals),
- CIA World Factbook <http://www.odci.gov/cia/publications/95fact/index.html>,
- Bartlett's Familiar Quotations - 1901 <http://www.columbia.edu/acis/bartleby/bartlett/>,
- Statistics Canada <http://www.statcan.ca/index.html>, or
  <gopher://gopher.statcan.ca/11/.English/Daily/Dailystat>,
- U.S. Patent Office <http://www.uspto.gov/web/menu/menu1.html>,

etc.

Good general access to secondary reference material on the Web is available from the second hierarchical levels of the Yahoo directory < http://www.yahoo.com/Reference/> and the Infoseek directory <http://home.netscape.com/home/internet-search.html> associated with the Infoseek search engine at this URL. Both these URLs lead to a wide variety of reference information including encyclopedias, dictionaries, on-line library systems, historical reference material, etc. Also, see the Web page maintained by the University of Delaware Library at <http://www.lib.udel.edu/ejrnls/>, which points to comprehensive listings of electronic journals. Online publishing is clearly the wave of the future, as current and historical academic and trade journals as well as news services such as magazines and newspapers will be made available online. Web links to this secondary material will continue to expand, and many of these services will recover their costs through advertising or direct charges to the searcher.

3.2 Information Search Strategies

There are several ways of classifying search strategies. We will use a classification with five categories (Meadow 1992; Cooper 1973). The strategy used will depend upon the objective of the search, and it will also depend upon whether information required is current (primary source) or widely accepted and well established information (secondary source), since this will determine the most likely information source. In the following, the five classifications are given, along with examples where the likely source would be either primary or secondary, or a combination of both, with the probable order in which they should be consulted. In one case, the information required is operational, relating to the implementation of the search process.

a) *Database exploration or browsing*

- widely accepted/established (e.g. information available *through* the Web: *secondary)*
- current (e.g. classes of information available *on* Web pages: *primary*)
- operational (e.g. how to go to previously visited Web pages: *Web browser-help*)

b) *General search*

- widely accepted/established (e.g. democracy*: secondary/primary*)
- current (e.g. tourist information*: primary/secondary*)

c) *Topic search*

- widely accepted/established (e.g. canine family animals: *secondary/primary*)
- current (e.g. availability/price of spreadsheet software*: primary*)

d) *Known-item search.*

- widely accepted/established (e.g. birthplace of Henrik Ibsen*: secondary*)
- current (e.g. date of the next Annual AIS Conference*: primary*)

e) *Exhaustive or existence search*

- widely accepted/established (e.g. all references to the War of 1812*.: secondary*)
- current (e.g. all information available on the ebola virus*: primary/secondary*)

Although this categorization is suitable for discussing potential search strategies, users do not necessarily commit to any particular strategy during the search process, and they may switch

from one type to another during a session (Meadow 1992).

4. Information Search on theWeb

There are three forms of meta-information (information about information) on the Web: a) Web page references, b) Web directories, and c) Web search engines. These are described in more detail in the following subsections. They point to almost all of the information appearing either on Web pages or accessible through Web pages, and are therefore the basis for developing Web search strategies. However, the existence of a reference to a Web page is not necessarily an indication that the page still exists. Web pages are maintained by thousands of independent entities, and pages may be altered, removed, added, or server addresses may change or servers may be removed without this information being made available to the information providers who reference the pages in question. Hence references to these pages may be out of date unless they are checked from time to time by those responsible for maintaining the references in other pages, directories, or search engine databases.

4.1 Web Page References

Almost all Web pages include references (URLs) to other related Web pages. In fact these references, taken together, constitute the logical Web network. From one Web page the searcher can simply click on any of the references and literally browse the Web in that topic area. This can be a highly inefficient way to gather information because there is no way to judge the relevance of the information to be found at that site without actually retrieving it, and it is very easy to lose sight of the object of the search and access much non-specific information while wandering around the Web in this manner. However, for expanding on concepts relating to a topic, this approach can be useful, provided the searcher has plenty of time available. This is also how Web crawlers automatically collect reference material (title and/or summary or introductory material from the first paragraphs of each page) which is then indexed in the databases available in the Web search engines. Most search engines also provide Web page forms for direct entry of information by Web page providers who wish to make sure their pages will be found by users.

4.2 Web Directories

Directories are useful if the searcher knows that the information of interest is likely to be in a particular topic area. The best-known Web directory is Yahoo (http://www.yahoo.com), which lists a large number of Web sites. Yahoo is organized hierarchically. It has 14 top level categories (Arts, Computers, etc.), and the searcher can "drill down" to lower and more detailed levels to find more specific sites which can then be accessed by clicking on their highlighted names. Although Yahoo can be useful to the uninitiated, there are other directories which are specifically oriented to particular topics, and are likely to be more complete than Yahoo in a particular topic area. A few examples are listed (not claimed to be complete!) in Appendix I, along with their URLs, for the topics: a) Business, Finance and Trade, b) Small Business and the Internet, and c) Distance Education. There is also a Clearinghouse for Subject-Oriented Internet Resource Guides, maintained by the University of Michigan, at <http://www.lib.umich.edu/chouse/chhome.html>. This lists directories in a variety of topic areas, and provides a quality rating for each directory listed. However, many recently developed Web directories do not appear in their listings, so the searcher who expects to make much use of recent Web information is advised to consult other sources as well (e.g. search engines) to find directories appropriate to the topic of interest.

4.3 Search Engines

Although directories are helpful in narrowing the search domain, for some topics the relevant information may fall in more than one domain. For example, there is a directory for neural networks, but for applications of neural networks in areas such as intelligent human-computer interfaces, one should also look in a directory for human-computer interfaces, or should do a specific search for this neural network application. And many specialized topics have no directory. Web search engines can be used in these cases, to search for relevant information. A Web search engine has an associated database which is indexed according to the occurrences of words appearing in Web pages it has indexed. Web pages are found by automatic "Web crawlers" which search the Web by tracking URL references from known pages to new pages, retrieving these pages and recording their contents for future reference by automatic text indexing

and abstraction into associated databases. There are a number of search engines available on the Web, accessible through their Web pages, but with forms that accept requests for key words or phrases the searcher wishes to use in searching the search engine's database. If one or more matches are found, summaries or titles of the pages are retrieved and displayed, complete with their URLs. If the page is of interest to the searcher, a simple click on the hypertext name of the page will access the indicated page through its URL, directly from its server. Some of these search engines index non-Web materials (e.g. usenet) and some provide access to interfaces to "secondary" materials such as library catalogs, encyclopedias, etc. not actually indexed by the search engine.

4.3.1 Search Engine Information Retrieval Modes

It should be emphasized at the outset that document indexing and search is unlike business database search. These differ because of the fundamental differences between data and documents (Blair 1990). Business data retrieval systems are said to be deterministic, while document retrieval systems are probabilistic. In a business database, the required record is either available and retrievable or it is not, by specifying a unique primary key or a non-unique field identifier. In document search, specifying particular words or phrases does not necessarily guarantee that all relevant documents indexed in the database will be retrieved. This is because there may be minor differences in notation between the search words and phrases and the indexed words and phrases, or entirely different words or phrases may have been used in the document to describe the information of interest to the searcher. An additional difference is that, because of the size and the dynamic nature of the Web database itself, none of the Web search engines have databases that even come close to indexing all available Web pages.

Web page references are retrieved from any of the search engine databases by entering key words or phrases into a form that provides user access to the search engine. Depending on the search engine being used, the user may also specify certain search control characteristics such as Boolean relations between the words/phrases, proximity to one another, position on the page (title, abstract, body, etc.), whether or not words in the retrieved page reference must match the

specified word exactly, etc. Page references retrieved are normally displayed in decreasing order of a score calculated for each document, which is based on the frequency of occurrence and/or the relationships or existence of all the words or phrases in the retrieval request. The following lists some of the control mechanisms the user may specify (depending of course upon the search engine being used). More detailed discussions are available elsewhere (Blair 1990; Pao 1989). In the following discussion, each entry within angle brackets represents one complete example for which a search may be requested (excluding the angle brackets) with a particular search engine. Note, however, that there is no standard input format for search engines, so these examples may not work in this exact format for any particular search engine.

Exact fit, or Simple search - The user specifies whole words or complete phrases that must appear within the indexed material (e.g. <*Cobol*>, <*Fortran Subroutines*>, or <*"Visual Basic"*>). In the second example, a reference will usually be retrieved if *Fortran* and/or *Subroutines* appear at any point within the referenced material (references containing both words will be ranked higher and hence listed first in order). In the third example, *Visual Basic* must appear as a phrase within the material if the reference is to be retrieved.

Boolean - Users may specify various existence relationships among key words and phrases using AND, OR, NOT, or equivalent terminology (e.g. <*Shakespeare* AND *play* BUT NOT *review*>).

Proximity - Users specify maximum distance relationships between key words and phrases, in terms of the number of words between them, and may also specify where the terms should be found, such as in the title, summary, or in the body of the page (e.g. <*project* WITHIN "summary" NEAR *evaluation*>).

Weighted Search - No search engine appears to use this technique (a previous version of Open Text did), but the following demonstrates how it would work. The user applies different weights to different key words or phrases, which could be used by the search engine in determining the value of the page references retrieved (e.g. < *"toy manufacturing"* WITHIN "title" WEIGHT 100, *"toy manufacturing"* WITHIN "summary" WEIGHT 20>).

Natural Language - Accepts the search request in terms of a natural language description

(Lewis & Sparck Jones 1996) . Although none of the search engines examined used natural language (possibly for reasons of efficiency), *Brittanica Online* provides an excellent example of a natural language interface (e.g. <*How can volcanic eruptions affect world climate?*> or <*What makes the sky blue?*>).

Fuzzy Search - Retrieves the page references where words may not exactly match key words provided, but there is some possibility of relevance/timeliness (e.g. <*tango* "Also search for related words"> . This is normally used in conjunction with a thesaurus of related terms or synonyms, and should not be taken to mean a combination of AND and OR for specified key words and phrases, as was suggested in the instructions for one of the search engines examined in this study.

Although there are currently a number of search engines available on the Web, this study reviewed only five: Open Text, Infoseek, Lycos, Excite, and Alta Vista. Table 1 summarizes their characteristics. All the search engines provided some form of ranking of the page references retrieved based on calculations that used the number of occurrences of key words or phrases in the referenced page, and references were listed in descending order of the ranking. All the search engines allowed the use of phrases. None of the search engines use weighted search or natural language, and none provided an effective use of fuzzy search, although Lycos used one form of fuzzy search in which users could specify word or phrase matches which ranged from "loose" to "strong". This could very helpful in narrowing a search to the most relevant pages. All the search engines except Excite allowed Boolean and/or proximity search. This imposed a severe limitation on the use of Excite - i.e. it could only be used effectively with single word or phrase searches. The response times from all the search engines depended, of course, on the time of day since use would be highest during the working day. However, after-hours search was extremely fast (no more than a few seconds for relatively complex searches) for all the search engines.

Search Attribute	Search Engine				
	Open Text	Infoseek	Lycos	Excite	Alta Vista
Boolean	Y	Y	Y		Y
Proximity	Y	Y	Y		Y
Phrases	Y	Y	Y	Y	Y
Ranks Results	Y	Y	Y	Y	Y
"Fuzzy Search"			Y 3)		

Notes: 1) No database size estimates are included, since the search engine vendors do not use a common indexing method and do not have a standard rating system. However, there seems to be little doubt that the Alta Vista search engine indexes the most Web pages among these five search engines.

2) Pages retrieved are displayed in declining order of "relevancy", calculated by algorithms specific to the search engines, based on key word occurrences.

3) Lycos allows users to specify matches, ranging from "loose" to "strong", relating to spelling or word fragments.

Search Engine URLs: Open Text: http://www.opentext.com/omw/f-omw.html
Infoseek: http://home.netscape.com/home/internet-search.html
Lycos: http://www.lycos.com/lycos-form.html
Excite: http://www.excite.com/
Alta Vista: http://altavista.digital.com/

Table 1

Some Search Engine Characterisics

Clearly, the search control mechanisms used and the size of the databases associated with these search engines will determine the number of relevant pages returned from a search, but the characteristics of the automatic text indexing used and the page titles and summaries retrieved during the searches are also important in determining the effectiveness of a search engine.

5. Search Engine Performance Effectiveness Measures

Given that a search and retrieval process has taken place on a Web search engine's database, there are four possible classifications of Web page retrieval outcomes (Blair 1990). The pages may be a) retrieved, relevant and timely (useful), b) retrieved but not relevant and timely (useless), c) not retrieved, but relevant and timely, or d) not retrieved and not relevant and timely. Given that it is difficult if not impossible to estimate the number of pages indexed in full by a search engine, the number of pages in class d) is not available for use in developing any outcome measures. Later in this report we will discuss how to estimate the number of relevant and timely documents available on the Web, but we know from other studies (Harman 1995) that searches of documents indexed from the same database by two different search engines yield only from 47% to 80% overlap between documents retrieved, even when the entire set of documents has been indexed by both search engines. This contrasts with the Web where no search engine fully indexes all its pages. In fact the Web appears to be outrunning the capability of existing search engines to keep up with its growth. A Web text indexing system is therefore very unlikely to provide 100% retrieval of relevant and timely documents on the Web, and the number of relevant, timely, but not retrieved pages (category c) above) can vary substantially, depending upon both the indexing mechanism used and the retrieval specification.

Web page relevance and timeliness is decided by the searcher at the time of retrieval. The question of completeness is also important but, as we have pointed out, the searcher who is willing to satisfice will not insist on completeness. We will address completeness when we discuss exhaustive search, where the searcher is not willing to satisfice but insists on retrieving all relevant and timely Web pages.

We will define the two most commonly used measures of retrieval, *Recall (R)*, and *Precision (P)*, (Meadow 1992) based on the number of Web pages found in each of the above categories (a,b,c,d, respectively). We will also include a relevance/timeliness interval scale for the Web pages, where relevance/timeliness s_i for page i can vary from 0 (not relevant/timely) to s_{max}

(Highly relevant/timely). A relevance/timeliness scale is important because it weights more relevant/timely pages more heavily and penalizes techniques that retrieve pages with little or no relevance/timeliness. In our study, we adopted an s_{max} of 1.0.

Recall *R*, (also Recall Ratio), is the ratio of the number of relevant/timely Web pages retrieved, (weighted by relevance/timeliness) to the number of relevant/timely Web pages in the database.

$$R = \frac{\sum_{i=1}^{a+c} s_i}{(a+c)\ s_{max}} \qquad [1]$$

Precision *P* is the proportion of retrieved pages that are relevant/timely, or the ratio of the number of retrieved pages, weighted by relevance/timeliness, to the total number of pages retrieved.

$$P = \frac{\sum_{i=1}^{a+b} s_i}{(a+b)\ s_{max}} \qquad [2]$$

The advantage of using Recall and Precision as measures is that they are not sensitive to the size of the collection. However, it is difficult if not impossible to determine the number of relevant/timely pages not retrieved (c) from the database, so this must be estimated indirectly. We will discuss a procedure for doing this in a following section.

There generally appears to be a relationship between *R* and *P*: high recall ($R \rightarrow 1$) implies low precision since this will probably include many irrelevant/non-timely pages among the pages retrieved ($P \rightarrow 0$), and high precision ($P \rightarrow 1$) implies low recall since many relevant/timely pages are likely to be missed but not many irrelevant/non-timely pages will be retrieved if the retrieval

conditions are very tight ($R \rightarrow 0$) (Meadow 1992). In fact both precision and recall depend upon how the retrieval was carried out and the relevance/timeliness values assigned to the retrieved pages. Attempts to develop overall retrieval effectiveness measures based on both R and P appear to have been relatively unsuccessful (Meadow 1992).

6. Example Exhaustive Search to Demonstrate Effectiveness Measures

In performing an exhaustive Web search on a particular topic, more than one search engine should be used. The difference in what is retrieved by several engines can also be very instructive in planning further searches, and it can help in estimating how much information on a particular topic is really available on the Web. To demonstrate the use of effectiveness measures and how several search engines can be used for an exhautive search on one topic, a search was carried out using the five search engines mentioned previously: Open Text, Infoseek, Lycos, Excite, and Alta Vista. Three levels of relevance/timeliness were used to evaluate retrieved Web pages: highly relevant/timely (1), somewhat relevant/timely (0.5), and not relevant/timely (0).

In the example used, it was supposed that we were interested in making home brewed beer, but before getting started we wanted to retrieve as much information as possible from the Web about different methods of home brewing. To keep the amount of information to a manageable level, the search was carried out for "wheat beer" only. Heeding the warning by Scoville (1996) that a search on this topic will give a lot of information on drinking wheat beer rather than brewing it, the topic was narrowed by searching on "wheat beer" & "recipe". All the engines display retrieved Web pages in descending order of an internally calculated score that has some relationship to the number of occurrences of the search keys, but they do not all use the same scoring mechanism. Consequently, some engines may retrieve many hundreds of Web pages, but only the few tens of pages presented first are likely to have a high degree of relevance/timeliness. To determine how many retrieved pages should actually be included in the count, pages were examined in order until it was clear that there were few if any remaining pages

that would rank as being highly relevant and timely to the topic. At that point, the count of that engine's retrievals stopped. Summaries of the pages as retrieved by each search engine from its database to that point were saved in a separate file for further analysis.

Table 2 summarizes the results of this study, with the number of pages retrieved by each engine, the number which coincided (overlapped) with pages retrieved by one or more of the other engines, and the percentage of overlap expressed as a ratio of overlapped to total retrieved. One finding was that the overlapped pages were generally found to be highly relevant and timely, agreeing with the results of another published study (Pao 1994). The relevance/timeliness frequency of the pages retrieved is also given in the table, and this was used to calculate the retrieval precision for each engine in this case. This varied from a high of 0.92 for Lycos to a low of 0.37 for Excite. The low precision for Excite is due almost entirely to the fact that this engine does not support Boolean search. Hence there were many non-relevant pages and non-timely mixed with the relevant and timely in the first tens of pages it retrieved. The total number of unique page occurrences among the five engines was 129, and when this was weighted according to their relevance/timeliness, the total was 84.

The "maximum estimated" recall for this particular type of page is also given for each engine, and this is simply the ratio of the number retrieved divided by the total unique (unweighted) occurrences. This varied from 0.13 for Open Text to 0.35 for Alta Vista. Note that these are maximum values because of the implicit assumption that there were only 129 such pages on the World Wide Web. This is unlikely to be the case, since there are probably Web pages that all five of the search engines missed. To account for these missing Web pages, Appendix I describes a model developed to assist in estimating the true number of such pages on the Web from the amount of overlap between retrievals by the different engines, under certain assumptions about the randomness with which the search engines gain access to the pages they are indexing.

The model was used with the "wheat beer & recipe" data and the results are shown in Table 3. Figure 1 is a plot of the probability distribution of overlaps predicted by the model of

Appendix I, for the estimated total number of Web pages (180) and the total number of pages actually retrieved by two search engines (30 and 38 respectively). The expected values predicted from this and similar distributions, with appropriate parameters applied, are given in Table 3 (with standard deviations in brackets) for each pair of search engine overlap/duplicate results, given the

Measure	Search Engine				
	Open Text	Infoseek	Lycos	Excite	Alta Vista
#Wheat beer & recipe	17	30	38	26	45
Overlap with Others	7	9	13	8	10
% Overlaps	41	30	34	31	22
Rel/Tim. Frequency	Open Text	Infoseek	Lycos	Excite	Alta Vista
Highly Rel/Tim. (1)	13	15	34	9	25
Somewhat Rel/Tim(.5)	0	8	2	1	9
Not Rel/Tim. (0)	4	7	2	16	11
Precision	0.76	0.63	0.92	0.37	0.66
Max. Est'd Recall	0.13	0.23	0.29	0.20	0.35
Total Unique Occurrences of wheat beer & recipe					129
Total Unique Occurrences Weighted by Relevance/Timeliness					84

Table 2

Analysis of Web Search Data on "wheat beer" and "recipe"

	Search Engine				
	Open Text	Infoseek	Lycos	Excite	Alta Vista
Relevant Pages	17	30	38	26	45
Overlaps (Open Text)		5	5	2	1
		2.83 (1.47)	3.59 (1.61)	2.46 (1.38)	4.25 (1.70)
Overlaps (Infoseek)			5	4	2
			6.33 (2.05)	4.33 (1.76)	7.50 (2.17)
Overlaps (Lycos)				3	5
				5.49 (1.93)	9.50 (2.38)
Overlaps (Excite)					3
					6.50 (2.05)

Notes: 1) There were a total of 129 unique Web pages retrieved by the five search engines on the topic 'Wheat beer' & 'recipe'.
2) Relevant Pages are the "total number of pages" (see definition in the text) retrieved on the topic by each search engine.
3) Overlaps (integer numbers shown) are the number of duplicate retrievals actually observed between the indicated pair of search engines.
4) Overlaps (decimal numbers shown) are the mean (with standard deviation in brackets) duplications estimated by the model, using $N = 180$. This value of N was selected to minimize the sum of squared differences between calculated means and observed values, excluding the Alta Vista data (see explanation in the text).

Table 3

Actual and Predicted Duplicate Retrievals Between Search Engine Pairs for Web Search on "wheat beer" & "recipe"

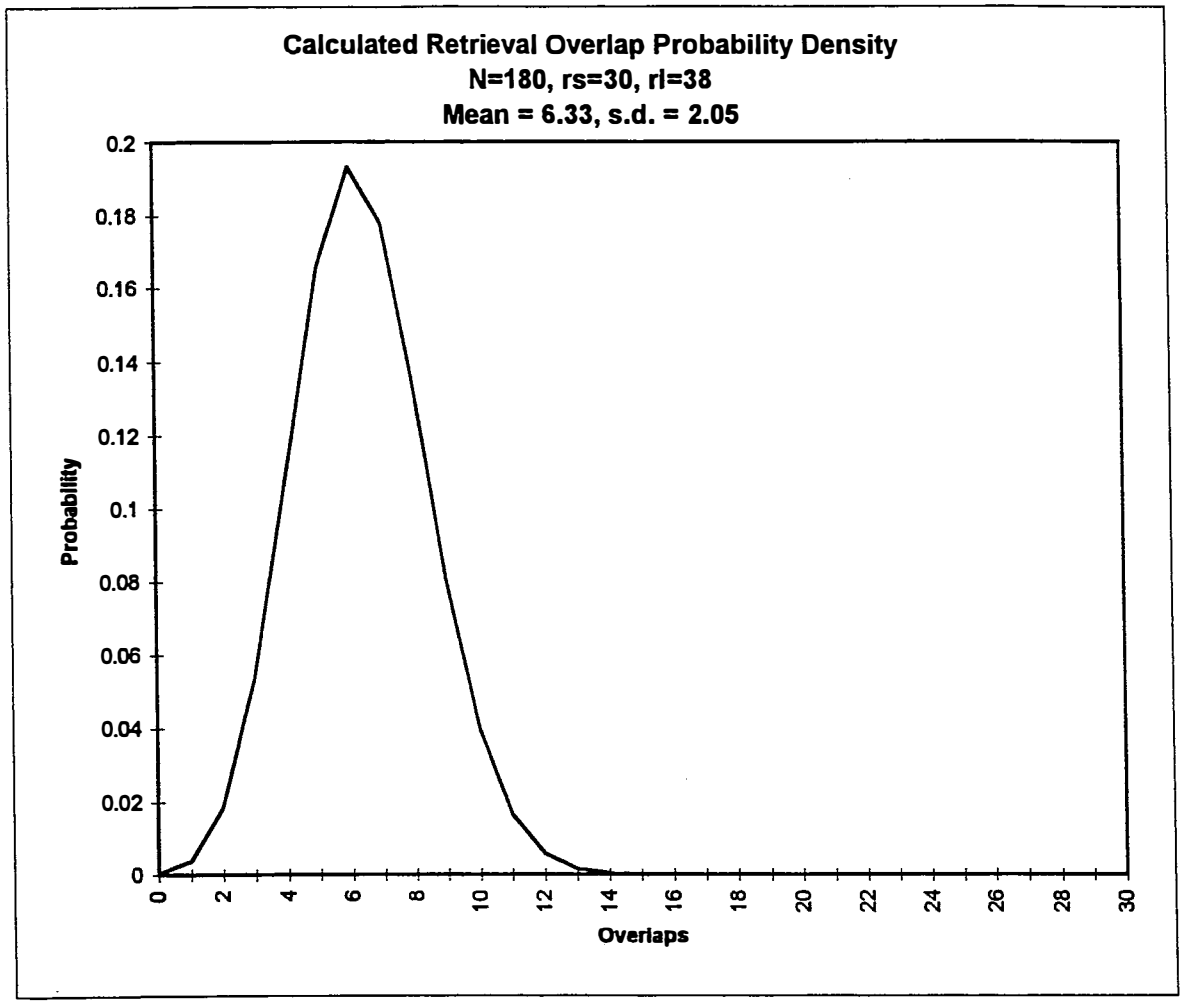

Figure 1

Calculated Retrieval Overlap Probability Density

total number retrieved by each of the search engines. The estimates are based on a total of N =180 total occurrences in the database. This number was chosen because it minimizes the sum of squared differences between the observed duplicates (also shown, above the expected values in each case) and the calculated overlaps over the first four search engines listed. Hence, this represents a best estimate of the total number of Web pages on this topic actually present on the Web. If this estimate is accurate, then the search was not truly exhaustive since only 129 or about 70% of the 180 relevant Web pages were retrieved among the five search engines used. Among individual search engines, Alta Vista retrieved the highest percentage (25%) of the estimated 180 relevant pages.

The Alta Vista data were not used in estimating N, because the number of overlaps between this search engine and the others is much lower than expected from the total numbers retrieved. The number retrieved by the Alta Vista engine also seems low, given the vendor's claim of a total database size which is very much larger than that of any of the other search engines. There are several possible explanations for these differences: a) this search engine is not sampling from the same Web pages as the other engines, b) it is using a much different indexing mechanism or different indexing terms, c) the mechanism that could be used to control page retrievals differed among the search engines, d) the assumption of randomness about the page selection and indexing mechanism may not be suitable, or e) a combination of the foregoing.

7. Recommended Search Strategies

Given the previous discussion and investigation, we have enough information at hand to develop some concrete recommendations for search strategies. These search strategies will clearly depend upon: a) the Search Objective (Exploration, General, Specific, Known-item, or Exhaustive), b) whether the information is likely to be current or well accepted and established fact, and c) whether or not it is anticipated that there will be a continuing need for search support on the topic of interest. Caution is advised in interpreting the strategies recommended here. Astute searchers will probably modify their strategies on the fly if their initial suppositions about probable information sources turn out to be incorrect. It is also important to keep in mind that directories are also referenced through search engines, so there is a great deal of cross-category availability of information. Since we cannot predict the future of Web information structures (for example, how the boundaries between primary and secondary information may change over time) these strategies may need revision as the Web develops and matures, and as supporting directories and search engines develop to meet the demand. Figure 2 is a graphical depiction of the proposed search strategies, discussed in more detail below.

Database Exploration or Browsing - this is basically a "get-acquainted" Web search, to determine what classes of information exist, and what forms this information is likely to take. The best

strategy in this case is to consult a directory such as Yahoo. These general directories have "Reference" at their top level, for following down to more detailed levels the information which is likely to be *secondary*. Otherwise the other categories of interest (e.g. "Arts", "Entertainment", "Science", etc.) can be examined for *primary* information sources.

Topic Search - In this type of search, there is a clearly identifiable topic, such as *Distance Education* using the *World Wide Web*, or *Shakespeare*'s play *Macbeth*. Only one of the search engines examined (Open Text) actually indexes every word on a Web page, so it would be possible (if desirable) to use this engine to retrieve all pages in its database which included the exact wording of a phrase such as *Distance Education using the World Wide Web.* In any case, it would be optimal to treat the first (*Distance Education* using the *World Wide Web*) as a primary topic and the latter *(Shakespeare*'s play *Macbeth)* as secondary, although the lines between primary and secondary blur in many cases, and both primary and secondary sources may contribute (e.g. if one wanted to find locations where Macbeth were currently being performed, one would use a primary search). A determining factor here would also be whether there is a continuing interest in this topic, or whether this is to be a one-shot search. If the former, then it is advisable to look for good directories first. If these exist, a lot of effort can be saved in the long run, since these will help to maintain continuing access to comprehensive sources of relevant/timely information. Otherwise, if the topic is primary, and the search is one-shot, it is advisable to begin by doing a primary search with a search engine.

When a search engine is used, it must be used in a controlled manner to ensure retrieval of mostly relevant/timely information. This determines the choice of search engine. Some allow very specific requests for key words or phrases through Boolean or proximity search operators, to help limit the number of useless Web page references retrieved. Search engines which do not have at least one of these features should be avoided in this situation. Secondary information is best found through reference listings of directories, whether the need is continuing or one-shot.

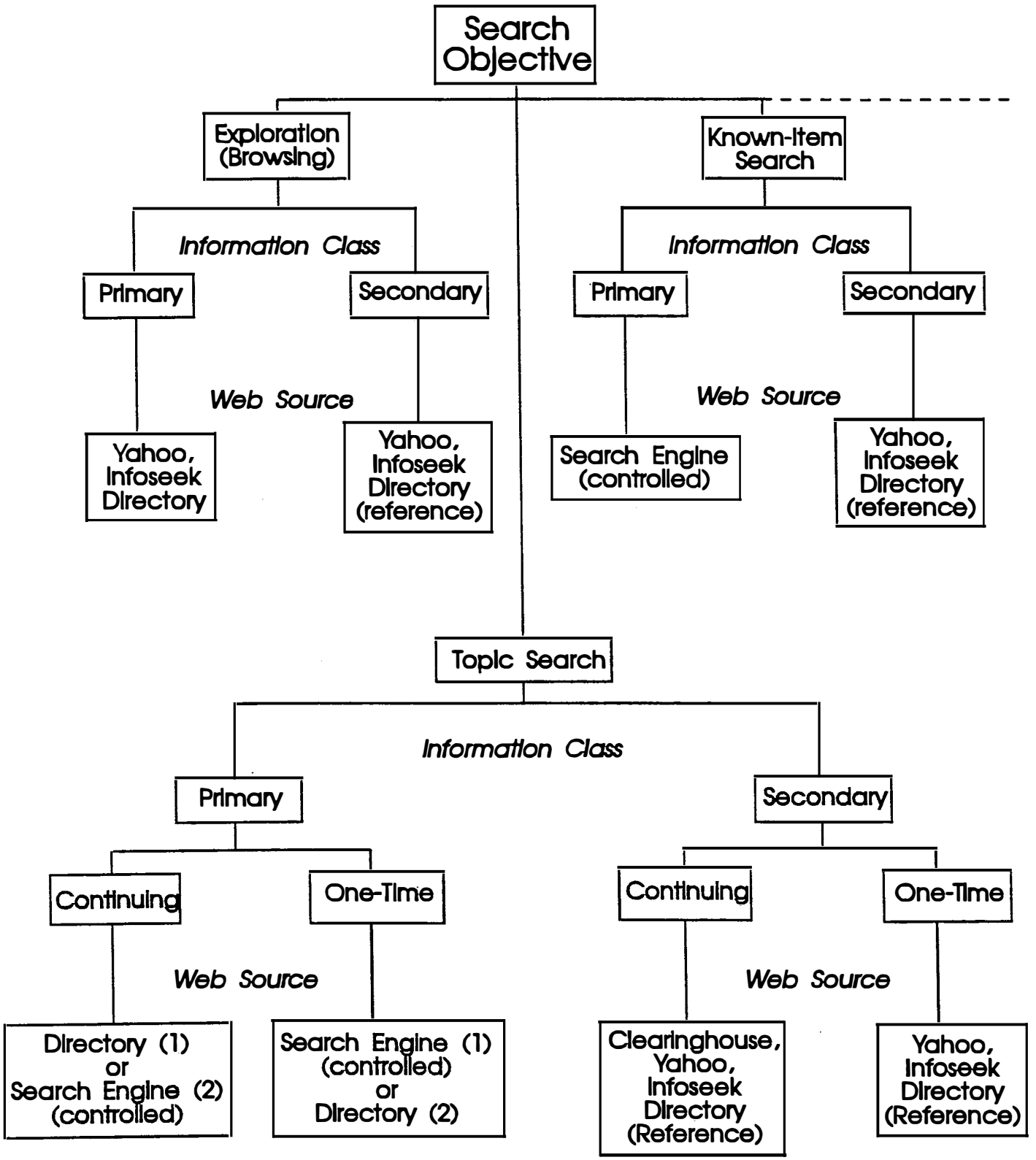

Figure 2

Web Search Strategies

Known-item Search - If the required information is likely to be primary, such as a conference date or recent news announcement, a tightly controlled search with one or more search engines until the required item is found will work best. However, there is no guarantee that the desired information will be in a page that is indexed in one or more search engine databases. Otherwise, if the information is secondary and the probable source is known in advance, a directory search is in order, to determine the availability and accessibility of an on-line secondary source (e.g. an encyclopedia).

General Search - The difference between a General search and Topic or Known-Item search is that a general search should be used if the searcher has only a few ideas or concepts about the topic of interest, and should therefore not restrict the search too tightly. The danger in a primary general topic search is that there will be too many totally irrelevant pages retrieved by the search engine if the topic is not well-defined (e.g. if a topic such as *primary education* is selected). If this is the case, the searcher can use an iterative or successive refinement approach to gradually narrow the search down by applying more restrictive controls, based on information seen or not seen in the pages retrieved in previous searches (if *primary education* is the general topic, then the searcher may narrow this to *science* in *primary education*, then narrowing this even further to *biology instruction* using *computers* in *primary education*, for example). For secondary information the searcher may have to check a number of directory references before coming up with a good source, and a search engine can be of potential help as well in this case.

Exhaustive Search - This was demonstrated for primary information in a foregoing section, and it involves the use of more than one search engine if resources are available, for a more comprehensive coverage of Web pages. It can also be used to provide an estimate of the total number of relevant/timely pages on the Web, to determine how effective the search has been. For secondary information, checking all the reference sources in the major directories is the best starting point. An exhaustive search would clearly be incomplete unless both secondary and primary sources were examined.

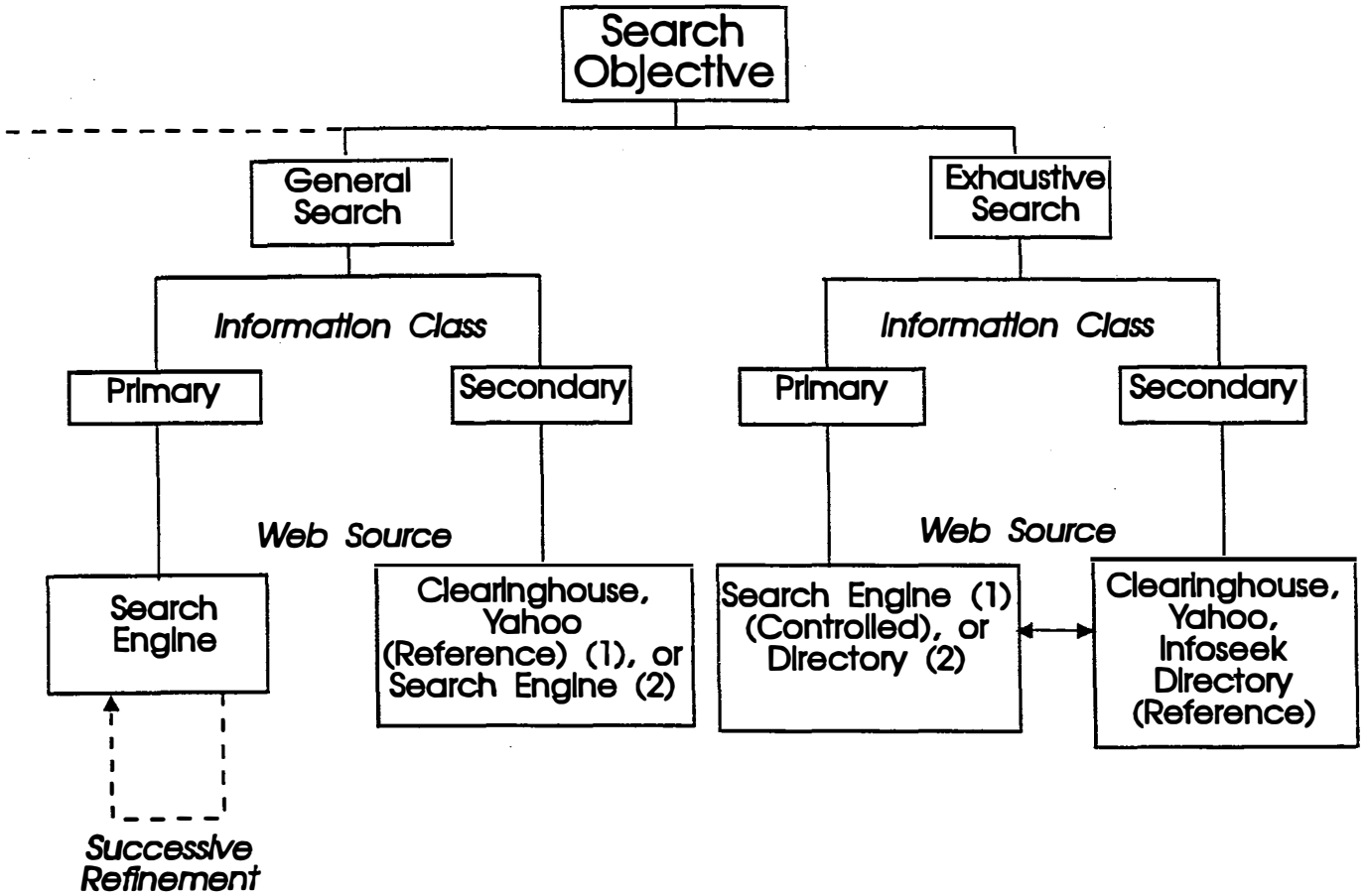

Figure 2 (continued)

Web Search Strategies

8. Discussion

In this paper we have attempted to provide a snapshot of the Web, a very new and dynamic source of information, and to suggest suitable strategies to retrieve relevant and timely information from its enormous and growing distributed database. Because of the enormity of this database, it is easy to waste large amounts of time browsing on the Web without focusing in on the topic of interest. Information search is often an ill-defined task in particular situations, such as when the objective is exploration (browsing) or carrying out a general search, and the searcher is likely to switch strategies during the search. In this case, the switch is likely to be to a more focused strategy such as topic search, as the searcher's interest becomes more specific. In the more focused information search cases: known-item search, topic search, and exhaustive search, the strategy can be well-defined in advance as we have shown. We have indicated how this type of search can be done most effectively, and we have also indicated appropriate beginning strategies for the less focused cases of browsing and general search.

What Of The Future?

The World Wide Web is a very new application of technology, and it is growing extremely fast to an overwhelming size, but many potential uses are not yet known because of supporting technologies yet to be developed (e.g. secure and verifiable transmission of payment for services or products, and wide availability of cheap, high speed links to home consumers, etc.). It would be presumptuous to try to predict the likely impact of the World Wide Web on the future of business and society. Nor would it be useful to speculate on what will be the most effective search strategies with tools which have not yet been envisioned for a system which is still evolving. However, it seems unlikely that any individual search engine will have the capacity to index all the information available on the Web or to handle the enormous demands placed on it by individual Web users. Some trends are beginning to appear:

1) at least one search engine's database is being made available on CD-ROM, so searches can be performed offline,

2) an integrated client utility is available from Quarterdeck called WebCompass© <http://arachnid.qdeck.com/qdeck/demosoft/webcompass_lite/> which simultaneously accesses multiple search engines and presents an integrated list of retrieved results. Since each search engine has its own specialized format for search controls, this may not work particularly well, especially when the number of retrievals from only one search engine is sometimes overwhelming,

3) another utility available (WebWhacker©) from Forefront Group Inc. allows downloading selected levels of pages and graphics from an entire site for local storage and off-line review at a later time (see <http://www.ffg.com/download.all.html>),

4) the quality and size of directories in specific areas of interest continues to improve, and this may well be the wave of the future for public Web use. What is lacking is a comprehensive and up-to-date directory of directories.

Concurrent with the growth of Internet sources, World Wide Web has also been adopted for local use, in business, government, and educational institutions which already have their own internal networks. This "Intranet" application normally denies access to external users in order to protect company private information. As an example, Digital Equipment Corporation has some 375 Web servers on its internal network, serving over 20,000 users. Other companies with extensive intranet installations include IBM, Sandia Laboratories, and Sun. Many universities are also exploring and experimenting with Web servers for providing education services that can be accessed by students either internally or externally from the university network. There are indications that intranet applications will grow to a greater level of importance than internet applications of the Web, and that the intranet will provide the glue that supports organizational memory through a diversity of primary and secondary information sources throughout the firm. Clearly, the ability to find relevant and timely information from this diversity of sources will be exceedingly important, and strategies must be developed for developing, maintaining, and using internal search engines, with indexes that record the locations of available information throughout the firm's intranet. This has the potential of significant contributions to the effective operations of every firm maintaining this type of information network.

References

Aronson, Larry (1994). *HTML Manual of Style*, Emeryville, CA: Ziff-Davis Press.

Berghel, Hal (1996). The client's side of the World-Wide Web, *Communications of the ACM*, 39(1), 30-40.

Berners-Lee, Tim, Cailliau, Robert, Luotonen, Ari, Nielsen, Henrik Frystyk, and Secret, Arthur (1994). The World-Wide Web, *Communications of the ACM*, 37(8), 76-82.

Blair, D.C. (1990). *Language And Representation In Information Retrieval*, New York, NY: Elsevier.

Catledge, Lara D., & Pitkow, James E. (1995). Characterizing browsing strategies in the World Wide Web, *Computer Networks and ISDN Systems*, 27, 1065-1073. (Also available at <http://www.gatech.edu/lcc/idt/Students/Catledge/browsing/user_1.html>)

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

Harman, Donna (1995). Overview of the second text retrieval conference (TREC-2), *Information Processing & Management*, 31(3), 271-289.

Lewis, David D., & Sparck Jones, Karen (1996). Natural language processing for information retrieval, *Communications of the ACM*, 39(1), 92-101.

Marchionini, Gary, & Shneiderman, Ben (1988). Finding facts vs. browsing knowledge in hypertext systems, *IEEE Computer*, 21(1), 70-80.

Meadow, Charles T. (1992). *Text Information Retrieval Systems*, San Diego, CA: Academic Press.

Pao, Miranda Lee (1989). *Concepts of Information Retrieval*, Englewood, CO: Libraries Unlimited.

Pao, Miranda Lee (1994). Relevance odds of retrieval overlaps from seven search fields, *Information Processing & Management*, 30(3), 305-314.

Scoville, Richard (1996). Find it on the Net, *PC World*, 14(1) (January), 125-130.

Senn, James A. (1990). *Information Systems in Management* (Fourth Ed.), Belmont, CA: Wadsworth.

Simon, Herbert A. (1960). *The New Science of Management Decision*, New York: Harper & Row.

Sundaram, Anita (1995). Towards the design of a hypermedia journal, *SIGOIS Bulletin*, 16(2), 23-25.

Appendix I

Some Examples of Topic Directories on the World Wide Web

Business, Finance and Trade

The Financial Data Finder (Dept. of Finance, Ohio State University)
<http://www.cob.ohio-state.edu/dept/fin/osudata.htm>

International Trade and Business (PACIFIC - Policy Analysis Computing and Information Facility in Commerce - EPAS Computing Facility at U. of T.)
<http://pacific.commerce.ubc.ca/trade/>

Bill's World of Business (Bill Henderson, bloorstreet.com web services, Toronto, Canada)
<http://www.bloorstreet.com/300block/bworld.htm>

Small Business and The Internet

Business and Economy: Small Business Information (Yahoo Directory - includes Internet Presence Providers, Credit Merchants/Electronic Merchant Systems/Transaction Clearing)
<http://www.yahoo.com/Business/Small_Business_Information>

Commerce and the World Wide Web (UCLA)
<http://www.deltanet.com/users/dplumley/eticket/contents.htm>

Distance Education

Canadian Directory of Distance Education Resources (Jason Merry, Dalhousie University - also includes Internet resources) <http://is.dal.ca/~jmerry/dist.htm>

The World Lecture Hall (University of Texas, Austin - Web delivery of course materials)
<http://www.utexas.edu/world/lecture/index.html>

Spectrum Virtual University (On-line adult education courses world-wide from a Southern California location) < http://horizons.org/ >

Appendix II

Duplications in Two Sets Drawn Randomly From the Same Population

Suppose we have a set Q composed of N discrete and distinctly identified objects. Suppose then that two subsets S and L are generated independently by selecting at random from Q as follows. First, r_s objects are selected without replacement from Q, and duplicates of these objects are stored in S. Then the objects are returned to Q, and r_l objects are selected randomly without replacement from Q and stored in L. We wish to determine how many of the objects now stored in S and L are identical.

For two non-negative integers a and b, with a ≥ b, the following combinatorial formula is defined in the usual manner:

$$\binom{a}{b} = a!/b!(a-b)!$$

Now, let $r_s \le r_l$, and let X be the random variable defined by the probability mass function $Pr(x)$, where x is the number of identical objects found in both S and L. Through elementary considerations of probability and combinatorial analysis the following expression can be developed in a straightforward manner, where $Pr(x=j)$ is given by:

No. of ways j common items can be selected from the N items in Q

× No. of ways the remaining $r_s - j$ items in set S can be selected from the $N - j$ items left in set Q

× No. of ways the remaining $r_l - j$ items in set L can be selected from the $N - j - (r_s - j)$ items left in set Q

÷ No. of all possible (and equally likely) outcomes.

Symbolically, this is represented by

$$Pr(x=j) = \frac{\binom{N}{j}\binom{N-j}{r_s-j}\binom{N-r_s}{r_l-j}}{\binom{N}{r_s}\binom{N}{r_l}} \qquad \max(0,\ r_s + r_l - N) \le j \le r_s \qquad [3]$$

Application

This formula can be used to estimate the size N of a Web sub-population on a particular topic, given the number of Web pages retrieved on that particular topic by two search engines. The approach is to find a value of N that gives an expected value E(X) calculated from equation [3] which matches the number of identical objects retrieved by the two search engines, given the total number of objects r_s and r_l retrieved on that topic by the two search engines, respectively. The value of N is, of course, at least as great as the number of unique pages retrieved by both the search engines combined, while r_s and r_l are the smaller and larger number of pages retrieved respectively by each search engine. The expected value or mean of the distribution given can be compared to the observed number of overlaps. If three or more search engines are used, then N should be selected so as to minimize the sum of squared differences between the observed overlaps and the expected number of overlaps across all pairs of search engines..

Caution is advised in interpreting results calculated in this manner, since certain assumptions may not hold: a) Web crawlers used to gather Web pages may not be sampling from the same database (e.g. the index of one may not have been updated for some time), b) the search engines may be using much different indexing schemes, c) the assumption of randomness about the page selection and indexing mechanisms may not be suitable, or d) a combination of the foregoing. Some search engines tend to index more pages in a cluster (from the same Web host) than do others, and this difference may also affect the randomness assumption.

Faculty of Business
McMaster University

WORKING PAPERS - RECENT RELEASES

395. Roy Adams, "A Pernicious Euphoria: 50 Years of Wagnerism in Canada", November, 1994.

396. Roy Adams, "The Determination of Trade Union Representativeness in Canada and the United States", November, 1994.

397. Shouhong Wang, Norman P. Archer, "A Fuzzy Decision Making Model", December, 1994.

398. Kalyan Moy Gupta and Ali R. Montazemi, "A Methodology for Evaluating the Retrieval Performance of Case-Based Reasoning Systems", December, 1994.

399. Jiang Chen and George Steiner, "Lot Streaming with Detached Setups in Three-Machine Flow Shops", December, 1994.

400. Jiang Chen and George Steiner, "Lot Streaming with Attached Setups in Three-Machine Flow Shops", December, 1994.

401. Ali R. Montazemi and Feng Wang, "An Empirical Investigation of CAI in Support of Mastery Learning", February, 1995.

402. Kalyan Moy Gupta and Ali Reza Montazemi, "Retrieval in Case-Based Reasoning Systems with Modified Cosine Matching Function", February, 1995.

403. Kalyan Moy Gupta and Ali Reza Montazemi, "A Connectionist Approach for Similarity Assessment in Case-Based Reasoning Systems", March, 1995.

404. John W. Medcof, "Selecting Alliance and Network Partners - Strategic Issues", March 1995.

405. Jiang Chen and George Steiner, "Discrete Lot Streaming in Two-Machine Flow Shops", March, 1995.

406. Harish C. Jain and S. Muthuchidambaram, "Strike Replacement Ban in Ontario and Its Relevance to U.S. Labor Law Reform", May, 1995.

407. Ali R. Montazemi and Kalyan Moy Gupta, "A Framework for Retrieval in Case-Based Reasoning Systems", June, 1995.

408. Ali R. Montazemi and Kalyan Moy Gupta, "An Adaptive Agent for Case Description in Diagnostic CBR Systems", June, 1995.

409. Roy J. Adams, Noel Cowell and Gangaram Singh, "The Making of Industrial Relations in the Commonwealth Caribbean", June, 1995.

410. Jiang Chen and George Steiner, "Approximation Methods for Discrete Lot Streamingin Flow Shops", June, 1995.

411. Harish C. Jain and S. Muthuchidambaram, "Bill 40 Amendments to Ontario Labour Relations Act: An Overview and Evaluation", June, 1995.

412. Jiang Chan and George Steiner, "Discrete Lot Streaming in Three-Machine Flow Shops", July, 1995.

413. J. Brimberg, A. Mehrez and G.O. Wesolowsky, "Allocation of Queueing Facilities Using a Minimax Criterion", January, 1996.

414. Isik Zeytinoglu and Jeanne Norris, "Global Diversity in Employment Relationships: A Typology of Flexible Employment", March, 1996.