

NOTICE

Warning Concerning Copyright Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be “used for any purpose other than private study, scholarship, or research.” If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of “fair use,” that user may be liable for copyright infringement.

This institution reserves the right to refuse to accept a copying order if, in its judgement, fulfillment of the order would involve violation of copyright law.

**SCANNED DOCUMENT
IS BEST COPY AVAILABLE**

Information Retrieval Systems

Tasks of information retrieval are accomplished in IR systems. Different types of information retrieval systems have been developed since the 1950s to meet different kinds of information needs. Online systems, CD-ROM systems, OPACs, and Internet retrieval systems are the four major categories of IR systems that have served users in various capacities to satisfy their information requests. In this chapter, each of the system types will be examined to show conceptually their features, functions, and capabilities in information retrieval.

8.1 Online Systems—Pioneer IR Systems

Online IR systems, also referred to as professional online systems, are often abbreviated as online systems or online databases. The word *database* is treated as a synonym for *system* in the latter case. They are the first kinds of IR systems that have applied computer technology. DIALOG and MEDLINE are two examples of online IR systems.

Online systems allow the user to search databases located remotely with the help of the computer and telecommunication technology. The systems initially only supported batch mode (i.e., an individual search request is not processed immediately after its submission but until a certain number of requests have been assembled), and later introduced real-time interaction between the user and system. There are three stages in the development of online systems (Bourne, 1980):

- Feasibility studies and demonstration projects—1950s
- Production with restricted user populations—1960s
- National or multinational IR services—1970s to present

Hahn (1996) presented a detailed account of the pioneers involved in the development of online systems. After the advancement and improvement made since the 1950s, online systems have become a distinct group of IR systems in the digital age with their own characteristics.

8.1.1 Features of Online IR Systems

Online systems collected mainly bibliographic plus some numeric information before the 1980s, and eventually included full-text information.

Multimedia information is not often seen in this type of retrieval system. Information included in online systems is selected and represented by professionals (e.g., indexers). Controlled vocabulary is extensively used for information representation and retrieval although keyword searching is supported at the same time. Command language was the norm for users to interact with online systems until menu selection was introduced in the 1980s. Graphical interface was gradually employed in the 1990s, especially when online systems started adopting the Web as the platform.

A great variety of retrieval techniques and IR models have been applied in online systems. Basic retrieval techniques, e.g., Boolean searching, case sensitivity searching, truncation, proximity searching, and field searching, are universally available in almost every online system. Advanced retrieval techniques such as weighted searching, fuzzy searching, and query expansion, though not supported in all online systems, can be found in some applications. In fact, online systems prior to the mid-1990s served as test-beds for IR research and development. They also functioned as showcases for new IR technologies before the emergence of Internet retrieval systems. For example, DIALOG implemented ranked output by introducing the RANK command in 1993 (Basch, 1993).

In addition, online systems can be regarded as a laboratory for acquiring information retrieval skills for several reasons. First, the system is constructed to allow structured practice and manipulation. Some online systems even set up special facilities (e.g., DIALOG's ONTAP files and workshops) to help train users. Second, online systems represented the *only* computerized IR systems at that time. Third, a *wide* range of retrieval skills can be tried and practiced in online systems. The fourth reason is nevertheless on the negative side: It is assumed that online system users must receive training before using it. The system is not intuitive and cannot be used simply by trial and error. As a result, end-users in online searching are generally information professionals who take up the role of search intermediaries between the system and the user with information needs.

8.1.2 Online Systems and Information Retrieval

Among the four types of IR systems identified in this chapter, online systems have the longest history and greatest influence in information retrieval. They implicitly set standards for others to follow. Whenever a new kind of IR system emerges, people always compare it with online systems to see how well the newcomer can do in the field. For example, Nahl-Jakobovits and Tenopir (1992) investigated the factors of response time, coverage, content, and cost in both the CD-ROM and online versions of Psychological Abstracts and Sociological Abstracts. Hildreth (1988) contrasted OPACs with online IR systems, and explored the incorporation of online system features in OPACs.

Chu (1998) compared Internet retrieval systems against online systems in aspects such as database structure, search capability, retrieval performance, output option, and user effort.

Online systems have apparently established themselves as the benchmarking systems in information retrieval. As mentioned previously, however, online systems are by no means perfect. High cost and user unfriendliness are attributes frequently ascribed to online systems. On the other hand, the pioneer role online systems have played in information retrieval is also well acknowledged.

8.2 CD-ROM Systems—A New Medium for IR Systems

CD-ROM systems emerged from the application of CD-ROM technology in information retrieval. CD-ROM systems are usually searched locally and do not rely on telecommunication for access if the systems are not networked. PsycLIT from SilverPlatter is one example of a CD-ROM system.

CD-ROM systems can be approximately regarded as online systems in the CD-ROM medium because these two kinds of IR systems share many features and because CD-ROM systems are modeled after online systems in many ways. Furthermore, CD-ROM systems were not implemented on a large scale until the 1980s when online systems had already become very influential in the field.

There seem to be no obvious phases in the development of CD-ROM systems. One main reason for this phenomenon is that CD-ROM systems were not an endeavor starting from scratch but rather a marriage between CD-ROMs and the mature online systems, using the former as a new storage medium. Therefore, the unique features of CD-ROM systems to a large extent are determined by the characteristics of the CD-ROM medium.

8.2.1 Features of CD-ROM Systems

Bibliographic, numeric, and full-text information remain the dominant information source processed in CD-ROM systems. But an increase in multimedia information storage occurred as CD-ROM technology became more capable of handling such types of information than online systems. Controlled vocabulary is used for information representation and retrieval in conjunction with natural language. Human involvement continues to be heavy in CD-ROM retrieval. Command language is seen less often in CD-ROM systems whereas menu selection is implemented more often. Graphic

interface made its debut in the CD-ROM environment, and some systems even applied the hyperstructure in their implementations.

Basic retrieval techniques (e.g., Boolean searching and proximity searching) are supported in CD-ROM systems while advanced search facilities (e.g., weighted searching and output ranking) are applied in a limited scope. It is not uncommon to see that one single system is stored on several CD-ROMs, which may restrict the application of some advanced retrieval technologies in the system.

Thanks to their high storage capacity and relatively small physical size, CD-ROM systems are theoretically portable because several disks can hold all the information needed in a database. However, the database alone cannot function as an IR system. Special equipment is required to run the system. On the other hand, it is more convenient for the user to search in CD-ROM systems, as they are not dependent on telecommunication technology to be functional. It is cheaper to conduct CD-ROM searches since they do not incur connection or other charges once the systems are purchased and installed. Users therefore can conduct searches without worrying if the clock is ticking and can concentrate more on the search itself.

The end-user, rather than the intermediary, does most of the searching in the CD-ROM environment. One reason is that CD-ROM systems usually impose no extra charges to the end-user according to the policy of fixed fee and unlimited access. Second, the interface of CD-ROM systems is friendlier than that of online systems as a result of the implementation of menu selection and graphic interfaces. Third, CD-ROM systems encourage browsing since online cost is not a concern to the end-user in this case. As discussed in Chapter 6, browsing and searching are the two major retrieval approaches while the former is favored in some circumstances. By comparison, browsing in an online retrieval session is costly.

CD-ROM systems, however, have limitations in updating. Each update means the replacement of existing CD-ROM disks, which significantly limits the updating frequency. Typical update frequency for CD-ROM systems is quarterly or bi-annually. Yet, the online counterparts of CD-ROM systems can be updated weekly, daily, or even continuously (e.g., Bridge World Markets News on DIALOG). Speed in searching CD-ROM systems can also be a problem, particularly when one system is spread over several CD-ROMs. CD-ROM systems also do not allow remote access if they are not networked.

8.2.2 CD-ROM Systems and Information Retrieval

CD-ROM systems have filled some voids that are not covered by online systems in information retrieval by reaching out to the end-user and by providing

retrieval services to more people for whom online systems might be too expensive or difficult to use. A large percentage of CD-ROM systems have corresponding online versions although their coverage and features may not be exactly the same.

CD-ROM as a storage medium is gradually being replaced by Digital Versatile Disk (DVD), a more recent optical disk technology. With two layers on each of its two sides, a DVD holds up to 17 gigabytes of video, audio, or other information. In comparison, a current CD-ROM disk of the same physical size holds less than 700 megabytes of information (Whatis.com, 2000). Although the replacement of CD-ROMs with DVDs has yet to take place in the field of information retrieval, the fact that DVDs represent a more advanced technology is likely to make CD-ROM an obsolete application. Will CD-ROM systems become extinct? Will CD-ROM IR systems be replaced by systems based on DVD or another emerging technology while the Web is used more and more often as the platform for information retrieval? Answers to these questions are yet to be found.

8.3 OPACs—Computerized Library Catalogs as IR Systems

Online Public Access Catalogs (OPACs) are traditional catalogs executed in a different medium (Malinconico, 1984). There were only some pioneering efforts, such as the Library Control System at the State University of Ohio, before 1980. Since 1980, however, prototype and operational OPACs have been installed in a steadily growing number of libraries (Hildreth, 1985). Few libraries nowadays do not have an OPAC in their systems.

OPACs, either stand-alone or as part of an integrated library automation system such as Horizon from AmeriTech, are an outcome of library automation. OPACs were first developed by both vendors and in-house teams. The latter option ceased to exist when library automation system vendors became more experienced and established. OPAC interfaces have been changed from command language to menu selection, form fill-in, and graphical presentations.

Hildreth (1984) classified OPACs into three generations to chart their recent history and predict their possible future design at that time. The first generation of OPACs emulates more or less the card catalog approach to file content, organization, and access for mainly known-item searching. The second generation of OPACs adds enhancements in subject access points, search capabilities, and other aspects. OPACs of this generation resemble online systems in many ways. The third generation of OPACs, becoming a reality nowadays, makes the division between OPACs and online/CD-ROM

systems less visible by having expanded access. The user can search online databases or CD-ROM resources via the OPAC of one's local library. Some other enhancements (e.g., integration of free text and controlled vocabulary search approaches) projected by Hildreth in 1984 for this generation of OPACs have also materialized.

8.3.1 Features of OPACs

Changes and developments are made as OPACs evolve from one generation to another. But there are certain features that seem unique to OPACs as one type of IR system. First, OPACs contain bibliographic information about library resources for institutions at various levels (e.g., local, regional, national). Although one OPAC is built for one library most of the time, there are occasions where one OPAC includes resources for multiple libraries. In contrast, the other three types of IR systems do not limit their coverage to bibliographical information at one institution or more.

Second, OPACs can be considered as an extension of MARC (i.e., a cataloging standard) records that are typically prepared by librarians using a set of rules and standards (e.g., classification schemes and cataloging rules) for the user population the library serves. Information representation in other kinds of IR systems cannot be done by using almost identical guidelines and with specific target users in mind. For example, DIALOG alone has hundreds of databases in its system. Each database is likely to be built and maintained using a different controlled vocabulary. While it is possible to define the user community of a library, it seems hard to clearly identify who would use an online system or CD-ROM system let alone the apparently immeasurable and ever-changing Internet retrieval systems.

Third, OPACs presently support at least field searching (e.g., author, title), keyword searching, and Boolean searching although library patrons make little use of the Boolean facility due to the reasons discussed in previous chapters. Known item searching constitutes a significant portion of all OPAC searches in part because OPACs are more likely to cover items "known" to the user than online or CD-ROM systems. Advanced search features (e.g., weighted searching) are usually not provided for OPACs since good retrieval performance is possible to achieve even by using the basic searching facility and thanks to the well-defined library collection.

Fourth, many present generation OPACs provide access to other resources including CD-ROM and online systems. With the help of Z39.50—a standard for retrieving information from different sources using a uniform interface—users can search other OPACs located at different places. Neither online nor CD-ROM systems can offer gateway services similar to OPACs mainly due to their for-profit nature.

Fifth, the design of OPACs encourages browsing in that OPAC users can typically browse by access points such as author and title while other types of IR systems usually present results chronologically or by relevance ranking. Browsing by call number in OPACs is particularly helpful to the user because the presentation order of the results bears a resemblance to the physical shelving order of library items. OPAC users thus are given the opportunity to locate more related items shelved next to each other before going to the shelves in person.

However, OPACs are still hard to use despite the fact that most of the OPACs now belong to the third generation. Borgman (1986; 1996) contributes the negative end result to the design of OPACs, which fails to incorporate lessons learned from information retrieval studies, and to insufficient understanding of searching behavior. The assumption that users without training should be able to use OPACs (Hildreth, 1988) is still not adequately supported today.

8.3.2 OPACs and Information Retrieval

The question about whether OPACs are library catalogs or online information retrieval systems was asked in the past (Hildreth, 1985). The answer to this question today should be that OPACs are information retrieval systems with their own characteristics. OPACs have been helping users locate materials in a library collection with functionalities unheard of when compared to their predecessors—card catalogs. In addition, OPACs also serve as a gateway to other IR systems by making linkages to them. As monographs are usually a major category of library materials, OPACs are the chief, if not the only, retrieval tool for accessing them. Without OPACs, library collections would become inaccessible in this digital age, as other kinds of IR systems are not designed specifically for that purpose.

8.4 Internet Retrieval Systems—The Newest Member in the Family of IR Systems

The exponential growth of the Internet makes it possible for people to access large quantities of digital information regardless of time and geographical locations. Meanwhile, it becomes obvious that additional IR systems are needed for retrieving any particular “needles” from the Internet “haystack.” Individuals, companies, and institutions have subsequently developed a great number of Internet retrieval systems, or IR systems for the Internet, to cope with the information explosion on the Internet.

As known, the phrase "Web search engines" seems a more popular name than Internet retrieval systems when people make reference to systems for retrieving information from the Internet. However, this author chooses to use the latter because the former appears too narrow in meaning. First, Web is only one application, although currently the dominant one, of the Internet. Information handled by other applications such as File Transfer Protocol (FTP) is also covered by Internet retrieval systems. Therefore, naming Internet retrieval systems as Web search engines implies that they only deal with information on the Web, excluding information originating from other Internet applications. Second, some Internet applications have their own search facilities (e.g., Archie for FTP information) although many have been phased out or started using the Web as the platform. Third, Internet retrieval tools can be directory-based (e.g., Yahoo!) or search-based (e.g., AltaVista). It would then be inappropriate to refer to directory-based Internet retrieval systems as Web search engines because searching was not supported in their original design. The searching component now available in many directory-based Internet retrieval systems is a result of the union between the two different types of systems, as will be discussed later.

More information will be presented in this book about Internet retrieval systems for two reasons. One is that Internet retrieval systems are newer compared with the other three types of IR systems. Another reason is that the other three types of IR systems have been studied in depth by many researchers over the years (e.g., online systems by Harter, 1986; Large, Tedd & Hartley, 1999; Meadow, Boyce & Kraft, 1999; Walker & Janes, 1999; CD-ROM systems by Chowdhury, 1999; Rowley & Slack, 1997; OPACs by Beaulieu & Borgman, 1996; Hildreth, 1985) while Internet retrieval systems have yet to be examined thoroughly and systematically.

8.4.1 Taxonomy of Internet Retrieval Systems

Internet retrieval systems, though with a short history, come in great numbers as well as varieties. It has been impossible to enumerate how many Internet retrieval systems are available on the Net since these systems began mushrooming in the mid-1990s. However, a categorization of Internet retrieval systems by the following criteria is intended to provide an overview of the newest member in the IR family.

8.4.1.1 By Retrieval Approach

Searching and browsing are the two major retrieval approaches described in Chapter 6. On the Internet, retrieval systems based on searching are called search engines (e.g., AltaVista and Google) whereas IR systems based on browsing are called directories (e.g., Yahoo!). Directories are also often

referred to as “catalogs.” Search engines let users compose their own search queries. By contrast, directories organize and present network resources under hierarchically structured categories. Users of directory-based Internet retrieval systems can locate information by following a predefined path, i.e., the hierarchy of categories developed for the system.

As discussed in Chapter 6, either searching or browsing as a retrieval approach has its limitations. For instance, search queries are difficult to formulate but required for conducting searches in search-based Internet retrieval systems. Internet users have to compose their own queries as information professionals usually do not act as intermediaries in this environment. Similarly, the number of results retrieved from the Internet for most topics can easily exceed thousands, if not millions. But browsing, as a retrieval approach, has no mechanism for narrowing a retrieval problem. Consequently, Internet retrieval systems attempt to surmount such difficulty by supporting both browsing and searching at the same site. The directory-based Internet retrieval systems license with search engines to provide the search component at their sites. The search-based Internet retrieval systems likewise contract with directory services to offer the browsing option at their sites. For example, Yahoo! used Google as its search engine and AltaVista was listed at its site directories compiled by Open Directory and LookSmart. The partnerships between these two kinds of Internet retrieval systems nevertheless change constantly. It would not be surprising to see, for instance, that Yahoo! changes to a different search engine for itself in the future as it did in the past. More specific information about how various search engines and directories form partnerships with each other for Internet retrieval can be found in the “Search Engine Alliance Chart” compiled by Danny Sullivan (Sullivan, 2001).

The combination of browsing and searching at one site produces the third type of Internet retrieval systems, namely, hybrid Internet retrieval systems. In the hybrid environment, the user can search or browse at the same site without switching to another. Retrieval effectiveness is improved subsequently.

8.4.1.2 By Application

Since the creation of the Internet, many different applications have been developed: Telnet, FTP, Gopher, Wide Area Information Servers (WAIS), and the Web or World Wide Web (WWW), to name some. Each of those applications performs certain functions. For example, Telnet is for remote login from a local system so that resources (e.g., hardware and software) at the remote site can be utilized. FTP is for transferring files between remote systems and local ones seamlessly at high speed. The Web is the largest information repository on the Internet because of its friendly user interface and hypermedia feature. Gopher,

an obsolete application today and completely superseded by Web, was designed for storing and retrieving resources on the Internet with hypermenu interface. WAIS, on the other hand, was a Z39.50-complied keyword search tool for Internet resources. Retrieval systems were built at one point or another for these applications except WAIS because it consisted of a retrieval mechanism itself. Table 8.1 lists retrieval systems created for major Internet applications under two different retrieval approaches.

Table 8.1 Major Internet Applications and Corresponding Retrieval Systems

Application	Search-based	Directory-based
Telnet	(Various)	Hytelnet
FTP	Archie	(None)
WWW	AltaVista, etc.	Yahoo!, etc.
Gopher	Veronica, Jughead	Gopher Jewels
WAIS	(Imbedded)	(Via Gopher)

As shown in Table 8.1, Internet retrieval systems comprise more than just Web search engines. Although many of the retrieval systems listed in the table no longer exist today, they did play a distinguished role in retrieving information from the Internet, particularly before the Web-based Internet retrieval systems were developed. To be more specific, Hytelnet provided a list of many Telnetable resources via a hypertext interface. It also allowed automatic logins for the sites listed. Archie, developed by Alan Emtage, a graduate student then at McGill University in Canada, periodically scanned anonymous FTP hosts and compiled information (e.g., host name, directory name, and file size) for files available on those hosts. People were able to search FTPable files by host name, directory name, and file name using Archie. Gopher enjoyed great popularity before the Web became the star application on the Internet. The retrieval systems designed for Gopher information—Veronica, Jughead, and Gopher Jewels—are rarely used as Gopher fades away from the Internet. The life span of WAIS was relatively short in part because of its command language interface and the emergence of Web-based Internet retrieval systems. However, the uniform interface as well as platform independent Z39.50 implemented in WAIS has been carried on in many other retrieval applications (e.g., OPACs).

Portions of the retrieval tasks previously accomplished by non-Web-based systems such as Archie have been transferred to Web-based systems. For example, Tile.net (<http://tile.net/ftp>) is a Web-based retrieval system devoted to finding FTP and other information. Other non-Web-based retrieval systems (e.g., Veronica) are not available anymore because their retrieval mission is fulfilled. Information from such Internet applications as USENET (for many-to-many discussions about numerous topics) and lists (for one-to-many discussions on selected subjects) is also retrievable via general (e.g., Google at <http://groups.google.com> for USENET information) or specific Web-based retrieval systems (e.g., Tile.net at <http://tile.net/lists> for list information).

In sum, Internet retrieval systems are at present all Web-based or use the Web as the platform. These retrieval systems include information coming from the Web and other applications (e.g., FTP and USENET) even though the Web is currently the largest information repository on the Internet.

8.4.1.3 By Content

Internet retrieval systems can also be categorized by the content of information they cover. Some of them maintain comprehensive coverage (e.g., AltaVista) by gathering information from various sources and subject areas. By contrast, others may be built for only one particular field. One such example is Gateway for Educational Materials (GEM at <http://www.geminfo.org>) for education.

Some Internet retrieval systems have evolved into portals by expanding the content coverage beyond searchable information. But search remains as the core in portals. Vortals, by comparison, are portals restricted to a vertical market (e.g., healthcare, insurance, automobiles, or food manufacturing). The market is vertical because it focuses on a relatively narrow range of goods and services, whereas a horizontal market is one that aims to produce a wide range of products and services (WhatIs.com, 2001). Roughly speaking, portals can be regarded as generic Internet retrieval systems. Vortals, on the other hand, are specialized in retrieving information targeted for one vertical market. Both portals and vortals are hybrid Internet retrieval systems that support browsing and searching.

In addition, there are retrieval systems devoted to retrieving special types of information on the Internet. For instance, SWITCHBOARD (<http://www.switchboard.com>) and FOUR11 (<http://www.four11.com>) of Yahoo! can be used for obtaining addresses and phone numbers. WHOWHERE (<http://www.whowhere.com>) of Lycos are created for locating e-mail addresses. MapBlast (<http://www.mapblast.com>) lets users type in an address and retrieve, for example, a detailed interactive map for that address.

This taxonomy of Internet retrieval systems illustrates the variety and capacity of IR systems developed over the years for the Internet. Internet retrieval systems, as a group, apparently have evolved and become the newest yet eminent member in the IR family.

8.4.2 Features of Internet Retrieval Systems

Internet retrieval systems, particularly Web search engines and directories, were not initially designed as part of the Internet development with the exception of WAIS. Instead, they were created as an afterthought when Internet users faced with an enormous amount of unprocessed information were without any tools for retrieval purposes. For example, Archie was developed when FTP users wanted to find out whether any anonymous FTP sites hosted the information they needed. Web search engines and directories were built after Web users found it increasingly difficult to locate what they wanted by just following hyperlinks imbedded at Web sites. If the entire Web had been treated as a library, we would have had the library collection before creating a library catalog for it. Besides, Internet retrieval systems differ from the other three types of IR systems in several aspects as discussed in the following sections.

8.4.2.1 Coverage and Source Information

Any information, before getting into online systems, CD-ROM systems, or OPACs, must go through editing, peer review, or similar checking procedures to ensure its quality. However, there are few quality control mechanisms in the information production process on the Internet. Anyone can put any information onto the Internet without its being checked for quality and suitability. Information of this kind would become the source information for Internet retrieval systems.

Internet retrieval systems rely primarily on automatic harvesting devices such as robots, spiders, or crawlers for gathering information from the Internet. These systems also allow people to submit information about their own sites although self-submission accounts for only a small percentage of all the information collected. Except for directory-based Internet retrieval systems, it appears rare to have human beings manually select information from the Internet. By comparison, information professionals are responsible for the data collection process in other kinds of retrieval environments.

When robots, spiders, or the like are dispatched for data collection, they are not all designed to copy the full content of a site. For example, Archie only copied the directory information of FTP sites. Some harvesting devices get the first 200 words or 20 lines of each site they visit. Others just obtain certain

information (e.g., title, headings, and hyperlinks) from them. Furthermore, Web sites as the largest information storage on the Internet have the networked hyperstructure. Few Internet retrieval systems have the intent and resources to collect information from every hyperlinked document. For example, some may decide to copy the documents by following one level of hyperlinks, leaving information at other levels untouched. As a result, most sites only have part of their full information included in Internet retrieval systems.

Besides the two factors just described, there is also the so-called “invisible Web” that affects the coverage of Internet retrieval systems. Invisible Web refers to those components that are not accessible by common Internet retrieval systems because of file formats (e.g., CGI scripts) or structures (e.g., databases mounted onto the Web). Portable Document Format (PDF) files on the Internet were not covered by any Internet retrieval system until Google changed its practice in early 2001.

According to studies conducted by Lawrence and Giles (1998; 1999), Internet retrieval system coverage relative to the estimated size of the publicly indexable Web (i.e., 800 million pages as of February 1999) has decreased substantially since December 1997, with no system indexing more than about 16 percent of the estimated size of the publicly indexable Web. In 1997, the coverage figure was about 33 percent. The conclusions were made after they examined six and 11 Internet retrieval systems respectively in the two studies using real queries performed by NEC Research Institute employees. Others (e.g., Smith, 2000) estimate that the top Internet retrieval systems fail to index 70–75 percent of the pages on the Web.

No matter what the actual coverage is, one thing becoming clear is that Internet retrieval systems are only able to process a decreasing portion of Internet information as the Net itself grows exponentially. The incomplete coverage of information and unselective practice of source information distinguish Internet retrieval systems from their counterparts in other IR environments.

8.4.2.2 Indexing Mechanism

Automatic indexing algorithms based on word frequency and similar criteria dominate the indexing practice in the creation of Internet retrieval systems. Limited human involvement is required in the indexing process even though some systems (e.g., Yahoo!) choose to manually categorize the information to be included in their databases. Controlled vocabulary is seldom adopted in indexing Internet materials in consideration of factors like cost, effectiveness, and the nature of information on the Net.

There are some projects (e.g., NetFirst at <http://www.oclc.org/oclc/netfirst/index.htm>) that apply controlled vocabularies for organizing

Internet resources. A list of similar projects can be found at CYBERSTACKS, a Web-based virtual directory of many undertakings germane to library and information science (McKiernan, 2001). But these projects, strictly speaking, function more as OPACs for Internet resources than as Internet retrieval systems. In addition, NetFirst is fee-based while the Internet retrieval systems under discussion are basically free.

Fields (e.g., author, title, and publication year) are identified and indexed when databases for online, CD-ROM, and OPAC systems are constructed. By comparison, Internet retrieval systems consist of no field data in the traditional sense in their databases other than the field-less index files for keywords and their corresponding locations in the system. Tagged information such as <title> ... </title> in HyperText Markup Language (HTML) or <price> ... </price> in eXtensible Markup Language (XML) can be treated at most as quasi-fields because they are not uniformly assigned during the indexing process.

In short, the indexing mechanism for Internet retrieval systems differs from that of other types of IR systems in that the former adopts the automatic approach based on keywords and does not index Internet resources by fields.

8.4.2.3 Searching Facilities

As one type of IR system, Internet retrieval systems provide virtually all the search facilities available in other retrieval environments. But there exist some points unique to Internet retrieval systems, which are discussed here.

Boolean searching, being taken for granted in non-Internet retrieval systems, is not supported in every Internet retrieval system. For example, the AND operator is not provided at Google as it is supposed to be since the order in which the terms are typed will affect the search results (Google, 2002; Rousseau, 2001). On the other hand, the plus (+) sign was used widely as a surrogate for the AND operator in the early years of Internet retrieval, which appears, however, misleading because the + sign represents the notation for the OR operator or the logical sum as explained in §7.2. Although that practice has been discontinued and the + sign now becomes the weighting symbol for weighted searching, it causes confusion to the end-user especially when the person has not done Boolean searching before.

Proximity searching is essential in retrieving phrases (if database information is not phrase indexed) and specifying the relative positions of search terms. Online systems are particularly strong in supporting proximity searches. By contrast, proximity searching has not yet become a universal feature for Internet retrieval systems. Among those systems that do support proximity searching on the Internet, few provide the full range of proximity operators that are typically available to online and other kinds of IR systems.

As for truncation, field searching, and case sensitive searching, Internet retrieval systems support them on a limited scale and with limited capabilities. Take truncation as an example, other IR systems can get very specific by, for instance, limiting the number of characters to be truncated. But most Internet retrieval systems can only go as far as unlimited right-hand truncation. Yet, automatic truncation is the default in some cases, generating more noise in search results. As mentioned in §8.4.2.2, there are only a few quasi-fields (e.g., title and URL) in Internet retrieval systems because such systems in effect are field-less. There are not many opportunities for the user to limit a search by field in Internet retrieval systems.

Weighted searching, not often seen in other kinds of retrieval systems, appears in the search facility repertoire of many Internet retrieval systems. The plus (+) sign is used consistently among Internet retrieval systems as the weighting symbol, meaning more weights for the term signaled. Based on the writer's current experience, weighted searching overrides Boolean searching on the Net. For example, in a query on *tax* AND *penalty* with a weight assigned to *penalty*, the results might contain only the term *penalty* while *tax* is totally ignored in the search.

Fuzzy searching, although not a common feature for other types of IR systems, does get implemented in some Internet retrieval systems. For instance, if one intends to search *Roosevelt Avenue* from Mapblast, an Internet retrieval system specialized in map information, a misspelled phrase of *Roosvelt Avenue* would still lead the searcher to the right location when other parts of the address (e.g., zip code) are correctly given. Google's "Did you mean: ..." feature, shown after a typo is made when entering a search query, is also an implementation of fuzzy searching.

Multiple database searching is not new, for example, to online systems. Databases suitable for performing such types of searching are usually set up or managed in one single system. For example, DIALOG is able to support multiple database searching by having several hundreds of databases that are similar structure-wise. In the case of Internet retrieval systems, cross database searching would be a more appropriate term to be used in this context because each singular retrieval system is constructed independently and different from others. They are not designed as subdivisions of a larger entity. Systems that can provide cross database searching on the Internet are called metasearch engines (e.g., Dogpile) or metaretrieval systems to be consistent with the terminology used in this book. The number of individual retrieval systems each metasearch engine covers varies from one system to another, normally in the range of half a dozen to two dozen. Usually single keyword searching works the best in cross database searching as there are more differences than similarities in terms of search syntax and semantics among

those individual IR systems. Some Internet metaretrieval systems remove duplicates before presenting results to the user.

Concept or meaning-based searching, rather than keyword matching, is particularly desirable on the Internet because it is crucial in improving retrieval performance. Yet, controlled vocabularies are not commonly used in Internet retrieval systems. Some attempts have been made in this respect. For example, SimpliFind (<http://www.simpli.com>) utilizes SimpliNet, a product of linguistics and cognitive science research, to automatically generate a list of concepts based on input query terms from which the user can choose. When a query is *book a flight*, for instance, SimpliFind would present a pull-down menu listing concepts such as *air travel*, *trip*, and *trajectory*. Then the user can choose a corresponding concept (e.g., *air travel*) to proceed with the actual search. Another example related to concept searching on the Net can be found at Oingo (<http://www.oingo.com>). Calling it a “meaning-based” searching engine, Oingo claims that it goes beyond searching for just simple text characters by allowing the user to refine a search based on the actual meaning of the query terms. Searches are conducted in Oingo’s semantic space, bringing up categories and documents that are close in meaning to the concepts the user is interested in. A particular word (e.g., *subway*) absent in a document would not exclude the possibility that the document is indeed conceptually relevant to the query on, for example, *public transportation*. Both SimpliFind and Oingo, strictly speaking, can only be considered as quasi-concept searching because their indexing does not appear concept-based—a prerequisite for concept searching. Rather, concepts are mapped onto keywords during searches using certain devices (e.g., SimpliNet). In the other three kinds of IR systems, controlled vocabulary is almost universally applied and concept searching is generally supported. SimpliFind and Oingo have now both changed from free to fee-based services.

Peer-to-peer (P2P) searching is a feature new in the IR field, which relates closely with the collaborative filtering approach used in commercial marketing. Search histories and bookmarks of previous users are examined when a new query is received to detect relevant information. When one joins a P2P network, that person’s query will be searched against the sites visited and bookmarked by other people belonging to the same initial network. This process can be repeated by expanding the search scope to additional P2P networks of which those people are also members. Festa (2001) describes P2P using Pandango, a project carried out by i5 Digital, as an example:

Pandango ... would determine relevance by examining a radiating network of “referrers.” Once someone downloaded Pandango and joined its peer-to-peer network, that person’s keyword search would examine the Web histories and bookmarks of an

initial network of 100 referrers. From there, the application would search the Web histories of those 100 referrers' combined 10,000 referrers, and once again – so that the query would canvass the Web pages visited and bookmarked by 1 million people. (p. 1)

P2P searching is distributed, making use of search information generated by previous users instead of searching in a centralized database. Certain issues (e.g., privacy and security) would inevitably arise as a result of P2P searching on the Internet. But they are beyond the scope of this book. P2P is also referred to as “bonding” in some publications (e.g., Schwartz, 2000). By comparison, neither online nor CD-ROM systems nor OPACs have the mechanism built in for conducting P2P searching.

8.4.2.4 Ranking Techniques

Ranking mechanism has been introduced to Internet retrieval as an effort to help the user locate “needles” in the huge “haystack” more effectively. Algorithms commonly employed for ranking search results in all IR environments include, as depicted in §5.1.2.2, term frequency, term proximity, term location, and inverse document frequency. While these algorithms rank search results according to the intrinsic attributes of documents (e.g., term frequency), the following three ranking methods focus on the extrinsic features generated by people when they access target Web documents or when they create their own sites.

The first new ranking technique, spearheaded by Google, is based on backlinks, i.e., links pointing to a site to be retrieved (Vidmar, 1999). Google assumes that the more times a site is pointed to by other sites, the more important it becomes. This practice appears very similar to citation analysis—an established method used for, among other things, evaluating the quality of scholarly publications. Borrowing some terminology from citation analysis, we can call the site being pointed to as the “cited site” while the site referring to the cited site becomes the “citing site.” In other words, Google has developed a mechanism for analyzing and ranking the links pointing to a cited site using PageRank, an algorithm named after one of its creators—Lawrence Page. Other parameters such as title, fonts, term proximity, and the importance of citing sites are also considered in the rank computation (Brin & Page, n.d.).

The second new ranking algorithm, known as the popularity approach, was developed by Direct Hit (<http://www.directhit.com>), which later became a part of Teoma (<http://www.teoma.com>). It ranks a site by the number of times users actually visit (thus the word “hit”). If we regard AltaVista and similar search engines as author-controlled IR systems (i.e., the relevance of search results is determined by how well keywords match with document content), if we call Yahoo! and similar directories editor-controlled IR systems

(i.e., editors locate and catalog sites by examining them one by one), Direct Hit represents a third kind of search mechanism: user-controlled IR systems in which search rankings depend on the choices made by users (Frauenfelder, 1998). The more times a site is visited by others, the more popular that site becomes, and the better ranking it receives.

The third ranking algorithm involves two steps. First, sites are ranked by using the traditional ranking methods such as frequency and proximity for analyzing hyperlinks contained within and the text around them. Second, these hyperlinks are further scrutinized using the backlink method Google employs. This ranking method, called "hypersearching" by its developers, is implemented in the Clever project at <http://www.almaden.ibm.com/cs/k53/clever.html> (Chakrabarti et al., 1999). Clever aims to identify the following two kinds of Web sites by hypersearching: 1) authorities – the best sources of information on a particular topic, and 2) hubs – collections of links to those locations. The reiteration of the above two steps locates and fine-tunes search results, which are eventually presented as pages of relevant links, separated into hubs and authorities. Hubs and authorities would receive higher ranking than regular sites. Clever is officially an IBM research project. So far it has offered only demo searches and research reports.

While traditional ranking methods may cause problems such as word stuffing (i.e., repeating some keywords deliberately in order to get a better ranking) or make little sense in situations like "pay for placement" (i.e., ranking search results based on the amount of fees paid), the three techniques described here open up a new dimension for ranking and evaluating Web sites. However, they may bring further biases into the ranking process as new and unlinked sites would have an increasingly difficult time becoming visible in those retrieval systems (Lawrence & Giles, 1999).

The backlink method adopted by Google is not novel in the online retrieval environment. The Institute of Scientific Information (ISI) has built a series of databases (e.g., Science Citation Index) since the 1960s according to the citation principle. But citation frequency was not implemented as a criterion for ranking search results in citation databases until the introduction of the RANK command by DIALOG in 1993. Direct Hit's popularity method and Clever's hypersearching algorithm, on the other hand, do not have counterparts in other retrieval environments where the ranking mechanism is still not widely available.

8.4.2.5 Search Modification

Modification of search queries or statements in online retrieval systems is well supported. The user can broaden or narrow down a search with various facilities (e.g., truncation or field searching) available in the system. CD-ROM systems operate in a manner similar to online systems because the former

can be treated to a certain extent as online systems in CD-ROM medium. OPACs, compared with online and CD-ROM systems, are less flexible in search modification. Since known item searching makes up a significant portion of all OPAC searches, the need for modifying searches is relatively small.

By contrast, Internet searches usually require modification because of the large amount of results each search produces and the difficulty in letting the end-user compose a precise query. Yet, it does not seem easy for users to modify their searches. First, no sets are created and stored for each search while this is a common feature in online and CD-ROM systems. Second, little modification mechanism is provided for refining searches. Lycos, for example, does permit users to conduct further searches within the results just retrieved by selecting the "Search these results" option. That is basically what the user can get in current Internet retrieval systems with regard to search modification. The flexibility and diversification of search modification supported in other retrieval environments are still absent on the Internet.

As controlled vocabularies are not commonly applied in processing network information, some systems (e.g., AltaVista) try to suggest a list of terms based on the query a user enters for modifying searches. For example, when a query on *tulips* is made, AltaVista provides the following terms for further refining the search: *Holland, plant, spring, tulip bulbs*, and more. The user can choose and decide which of the terms in the listing to be included in the search. Even with such features for search modification, it remains extremely hard for Internet users to narrow down a search to a desired level. Relevance feedback, the automatic approach for expanding a search, is, however, available in systems such as Google. The user can do relevance feedback by simply clicking on hyperlinks labeled as "Similar pages" or "More like this," for example. The concern is how useful this approach is when the user seems already overwhelmed by the number of search results the system retrieved even without using the relevance feedback option.

8.4.2.6 Interface

Command language and menu selection are used as the interface in a large percentage of online, CD-ROM, and OPAC systems. Internet retrieval systems created for applications other than the Web (e.g., Archie and Veronica) had interfaces of similar kinds. As the Web becomes a major Internet application and popular platform for other applications (e.g., USENET), Internet retrieval systems tend to all have the Web interface, which is graphical and hyperstructured. The uniformity in interface contributes to the user friendliness of Internet retrieval systems as a whole, a characteristic hard to attain by other kinds of IR systems.

In summary, Internet retrieval systems have unique features in coverage, source information, indexing, searching facilities, ranking methods, search

modification, and interface when compared with online, CD-ROM, and OPAC systems. The searching of multilingual and multimedia information should also differentiate Internet retrieval systems from other types of IR systems. That topic is covered in Chapter 9, Retrieval of Information Unique in Content and Format.

8.4.3 Generations of Internet Retrieval Systems

Internet retrieval systems have grown into a new member of the IR family. A lot of changes and enhancements have been introduced to them since their inception. Although having only a short history of development, Internet retrieval systems can be classified into three different generations based on the attributes presented in Table 8.2.

The first generation of Internet retrieval systems refers to such pioneer systems as Yahoo! and WebCrawler. Text information was the major type of information those systems collected. No particular attention was paid to multimedia information processing when this generation of Internet retrieval systems was built besides displaying the images already embedded on the pages or sites along with the text information retrieved. Directories and search engines were separated, and each system provided one retrieval approach, either browsing or searching. These systems did keyword indexing and supported, as one generation, basic searching capabilities (e.g., Boolean search, truncation, and proximity search). No search modification was possible with this generation of systems. Search outputs were not ranked but users were given the opportunity to choose formats (e.g., title only or title with a brief summary) and the number of results (e.g., 10, 20, or 30) a system would present without refreshing the screen.

Table 8.2 Three Generations of Internet Retrieval Systems

Attribute \ Generation	1st Generation	2nd Generation	3rd Generation
Information Covered	Mainly text	Text or nontext	Multimedia
Retrieval Approach	Browsing & searching separated	Browsing & searching combined	Browsing & searching integrated
Indexing	Keyword	Keyword with concept mapping	Keyword & concept
Searching Capability	Basic	Basic & advanced	Precision improving
Search Modification	No	Limited support	More narrow-down features
Output	Not ranked	Ranked	Personalized

The second generation of Internet retrieval systems grows significantly in number. Some systems represent improvements over their initial versions (e.g., Yahoo! and AltaVista) whereas others are new additions (e.g., Google and Teoma). A considerable number of the second-generation systems not only provides text information retrieval services but also helps locating image, video, and MP3 files. The retrieval process is, however, limited to databases or files that contain a particular type of multimedia information. In other words, text and multimedia information is processed separately in this generation of Internet retrieval systems. Although powered by different companies, both the browsing and searching mechanisms are presented at the same site so that the user can make a choice of retrieval approaches without switching to another site. Indexing is still based on keywords but some efforts have been made to map concepts onto keywords using devices such as SimpliNet from Simpli. While basic searching continues being supported by the second-generation Internet retrieval systems, advanced search facilities (e.g., weighted search) are introduced. Most Internet retrieval systems classify their searching capabilities into two kinds: 1) searches free of symbols or notations (e.g., operators) as simple searching, and 2) queries with symbols and notations as advanced searching. Search forms are normally used for advanced searches. Searches can be broadened and narrowed down in a limited number of ways, such as by date, by language, search within results, or relevance feedback. Search results are ranked most of the time based on a ranking algorithm whose composition is usually kept secret. Choices of output number and display format gradually fade away in this generation of Internet retrieval systems due to their diminishing significance to the user. As IR systems, the second generation has made a lot of progress in functionality and performance.

The third generation of Internet retrieval systems should have further enhancements in attributes listed in Table 8.2. First, multimedia will not be treated separately from text information in the new generation of Internet retrieval systems. Rather, techniques for processing and retrieving multimedia will be integrated with those for text information. Since the Web is noted for its capability in presenting multimedia information and has become the major Internet application, there is no reason why multimedia cannot be represented, processed, and retrieved together with the text information as research in multimedia IR advances. Second, directories and search engines are being integrated rather than simply offered at the same site. The user can search within a browsable category (e.g., Yahoo!) and search results are grouped into categories for browsing (e.g., Teoma). In addition to keyword indexing, concept indexing will be done when further progress has been made in research on natural language processing and representation. Search capabilities supported by previous generations will be improved and

enhanced with a focus on improving precision. Additional search facilities (e.g., concept searching, natural language searching) will be brought into the repertoire of retrieval capability the system maintains. The search modification mechanism should be oriented more toward effectively narrowing down searches because the user is usually overwhelmed by the huge numbers of results produced by Internet retrieval systems. Search result presentation will ideally be ranked and customized according to the user's specification instead of the system's predefined options. For instance, Fast (<http://www.alltheweb.com>) in its advanced search interface allows the user to decide, among other things, if offensive results should be eliminated or if pages whose sizes are larger than certain kilobytes should be presented. The third generation of Internet retrieval systems should be able to show noticeable enhancement of retrieval performance.

As seen from these descriptions, Internet retrieval systems have apparently passed the first generation, and are evolving well into the second generation. The third generation is emerging on the horizon with great potential. In less than a decade of time, we have experienced the change of generations in Internet retrieval systems because of the rapid development of the Internet for which those retrieval tools have been built. Their significance in the IR field and especially in the digital age cannot be overstated.

8.4.4 Internet Retrieval Systems and Information Retrieval

The Internet has become not only a gigantic information warehouse but also a popular platform for accessing other kinds of IR systems. Internet retrieval systems no doubt are the only tools for retrieving information from the Internet. Meanwhile, they also function as gateways to other information sources. As more and more information is put onto the Internet, the importance of Internet retrieval systems in the field will increase in proportion.

In addition, Internet retrieval systems are turning into the lab and showcase for new, advanced, and sophisticated IR techniques, a role that online systems have played in the past. To stay current with the most recent developments in the IR field, one should closely monitor Internet retrieval systems instead of online or other kinds of systems. Changes occur constantly, if not on a daily basis, in Internet retrieval systems. Net-based newsletters and sites (e.g., Notess, 2002; Sullivan, 2002) are created to inform the user of the latest developments about them.

On the other hand, Internet retrieval systems as a group are notorious for their low precision search performance. Information retrieved from the Internet may not be accessible due to factors such as broken and dated links, a consequence much less likely in other IR environments. The user must also

be responsible for judging the quality of information since the quality control mechanism for information on the Internet, if available, is at best very loosely structured (e.g., at the time of cataloging Internet resources for directory-based systems).

Searching, according to Brewer (2001), is the most visible and important aspect of the Internet after communication. Internet retrieval systems are designed specifically for retrieval purposes. Moreover, these systems are free of charge to end-users if they already have Internet access. It is therefore not an overstatement to say that Internet retrieval systems hold a prominent place in the field of information retrieval at present and will do so for a long time to come.

8.5 Convergence of Various IR Systems

It can be seen from the descriptions and discussions in this chapter that each kind of IR system has its features, functions, and capabilities in the area of information retrieval. Yet they are all designed and maintained for information retrieval purpose. In order to provide better and more convenient IR services to the user, the convergence of various IR systems has occurred in recent years. Convergence may take any of the following forms.

The first form of convergence, chronologically speaking, is the loading of CD-ROM systems onto OPACs. Since OPACs are generally created for individual libraries that also house CD-ROM systems, many institutions consolidate these two systems. Links are made from the OPAC to CD-ROMs, allowing the user to access two different IR systems at one place. Some institutions also establish links from their OPACs to online systems so that the connection procedure can be eliminated when the user conducts any online searches. This form of convergence takes place basically at the system connection level.

The second form of convergence occurs at the system database content level between OPACs and online systems. For example, an OPAC search reveals that a particular item is not in the library collection. The query for that item can then be submitted to the online system that the OPAC is linked with to see if the item is in the online system. Document delivery systems (e.g., ProQuest or UnCover) would very likely be used to obtain a physical copy of the item. But it is the linkage between OPACs and online systems that truly facilitates the retrieval process.

The third type of convergence happens when the hyperstructured Web has increasingly been used as the platform for Internet retrieval systems as well as online, CD-ROM, and OPAC systems. These IR systems become so interrelated that the boundaries among them are getting fuzzy. For example,

Field 856 of MARC, a cataloging standard, is allocated for recording URLs of Internet resources. The end products of this practice by nature would be, as indicated in §8.4.2.2, OPACs for Internet resources. In addition, more and more books and journals are being published electronically on the Web, integrated into the library collection, and represented in the OPACs. Therefore, information on the Internet, previously the unquestionable target of Internet retrieval systems, is included in OPACs that have been reserved exclusively for representing and retrieving library collections in the past. This practice has brought about the convergence of OPACs and Internet retrieval systems. Another example of such convergence can be found where Web documents list references or citations that are covered by online systems. Assuming a Web document along with its references is located with the help of an Internet retrieval system, the user can further search in the online system containing the hyperlinked references. The association between online systems and Internet retrieval systems is seamlessly established.

From a broader perspective, the Web is becoming the platform for using all four different types of IR systems. From a narrower viewpoint, it is not unthinkable for a user to have access to a Web-based library information system that includes all four kinds of IR systems, namely, the library's OPAC, online and CD-ROM systems the library subscribes to, and some Internet retrieval systems the library chooses to present. Will the Web become the only platform for IR systems in the future? Will a common platform catalyze further convergence among the IR systems? What will the impact of the convergence be on the user? We do not yet know the answers.

References

- Basch, Reva. (1993). Dialog's Rank command: Building and mining the data mountain. *Online*, 17(4), 28-35.
- Beaulieu, Micheline, and Borgman, Christine L. (Guest editors). (1996). Special topic issue: Current research in Online Public Access Systems. *Journal of the American Society for Information Science*, 47(7), 491-583.
- Borgman, Christine L. (1986). Why are online catalogs hard to use? Lessons learned from information retrieval studies. *Journal of the American Society for Information Science*, 37, 387-397.
- Borgman, Christine L. (1996). Why are online catalogs still hard to use? *Journal of the American Society for Information Science*, 47(7), 493-503.

- Bourne, Charles P. (1980). On-line systems: History, technology, and economics. *Journal of the American Society for Information Science*, 31(3), 155–160.
- Brewer, Eric A. (2001). When everything is searchable. *Communications of the ACM*, 44(3), 53–55.
- Brin, Sergey, and Page, Lawrence. [No publication year]. The anatomy of a large-scale hypertextual Web search engine. <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>. (March 18, 2001)
- Chakrabarti, Soumen, et al. (June 1999). Hypersearching the Web. *Scientific American*. <http://www.sciam.com/1999/0699issue/0699raghavan.html>. (May 26, 1999)
- Chowdhury, Gobinda G. (1999). *Introduction to Modern Information Retrieval*. London: Library Association Publishing.
- Chu, Heting. (1998). Internet search services vs. online database services. In Martha E. Williams (Ed.), *Proceedings of the 19th National Online Meeting*, (pp. 69–75). Medford, NJ: Information Today.
- Festa, Paul. (February 26, 2001). Search project prepares to challenge Google. *CNET News.com*. <http://news.cnet.com/news/0-1005-202-4950537.html>. (February 26, 2001)
- Frauenfelder, Mark. (September 25, 1998). The future of search engines. *The Industry Standard: The Newsmagazine of the Internet Economy*. http://www.thestandard.com/articles/article_print/0,1454,1826,00.html. (May 26, 1999)
- Google. (2002). The basics of Google search: Automatic “and” queries. <http://www.google.com/help/basics.html>. (August 29, 2002)
- Hahn, Trudi Bellardo. (1996). Pioneers of the online age. *Information Processing & Management*. 32(1), 33–48.
- Harter, Stephen P. (1986). *Online information retrieval: Concepts, principles, and techniques*. New York: Academic Press.
- Hildreth, Charles R. (1984). Pursuing the ideal: Generations of online catalogs. In B. Aveney, and B. Butler (Eds.), *Online catalogs, online reference*:

- Converging trends. *Proceedings of a Library and Information Technology Association Preconference Institute* (pp. 31–56). Chicago: American Library Association.
- Hildreth, Charles R. (1985). Online public access catalogs. *Annual Review of Information Science and Technology*, 20, 233–285.
- Hildreth, Charles R. (1988). Online library catalogues as information retrieval systems: What can we learn from research? In P.A. Yates-Mercer (Ed.), *Future trends in information science and technology. Proceedings of the Silver Jubilee Conference of the City University's Department of Information Science* (pp. 9–25). London: Taylor Graham.
- Large, Andrew, Tedd, Lucy A., and Hartley, R.J. (1999). *Information seeking in the online age: Principles and practice*. London: Bowker-Saur.
- Lawrence, Steve, and Giles, C. Lee. (April 3, 1998). Searching the World Wide Web. *Science*, 280(5360), 98–100.
- Lawrence, Steve, and Giles, C. Lee. (July 8, 1999). Accessibility of information on the Web. *Nature*, 400, 107–109.
- Malinconico, S. Michael. (1984). Catalogs & cataloging: Innocent pleasures and enduring controversies. *Library Journal*, 109(11), 1210–1213.
- McKiernan, Gerry. (2001). Beyond bookmarks: Schemes for organizing the Web. <http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm>. (8/14/02)
- Meadow, Charles T., Boyce, Bert R., and Kraft, Donald H. (1999). *Text information retrieval systems*. Orlando, FL: Academic Press.
- Nahl-Jakobovits, Diane, and Tenopir, Carol. (1992). Databases online and on CD-ROMs: How do they differ? Let us count the ways. *Database*, 15(1), 42–50.
- Notess, Greg R. (2002). Search Engine Showdown: The Users' Guide to Web Searching. <http://www.notess.com>. (August 21, 2002)
- Rousseau, Ronald. (2001). Google search. [List messages]. sigmatrics@listserv.utm.edu. (January 7, 2001 & January 9, 2001)

- Rowley, Jennifer, and Slack, Frances. (1997). The evaluation of interface design on CD-ROMs. *Online and CDROM Review*, 21(1), 3-11.
- Schwartz, Candy. (2000). Meeting review: Notes from the Boston 2000 Search Engine Meeting. *Bulletin of the American Society for Information Science*, 26(6), 26-28.
- Smith, Ian. (2000). The invisible Web: Where search engines fear to go. <http://www.powerhomebiz.com/vol25/invisible.htm>. (December 12, 2000)
- Sullivan, Danny. (2001). Search engine alliance chart. <http://searchenginewatch.com/reports/alliances.html>. (August 29, 2002)
- Sullivan, Danny. (2002). Search Engine Watch. <http://searchenginewatch.com>. (August 29, 2002)
- Vidmar, Dale J. (1999). Darwin on the Web: The evolution of search tools. *Computers in Libraries*, 19(5), 22-28.
- Walker, Geraldene, and Janes, Joseph. (1999). *Online retrieval: A dialogue of theory and practice*. 2nd ed. Englewood, CO: Libraries Unlimited.
- Whatis.com. (2000). Digital versatile disk. Entry updated on August 20, 2000. (March 4, 2001)
- Whatis.com. (2001). Vortal. Entry updated on January 26, 2001. (February 25, 2001)