Information Retrieval on the Web

MEI KOBAYASHI and KOICHI TAKEDA

IBM Research

In this paper we review studies of the growth of the Internet and technologies that are useful for information search and retrieval on the Web. We present data on the Internet from several different sources, e.g., current as well as projected number of users, hosts, and Web sites. Although numerical figures vary, overall trends cited by the sources are consistent and point to exponential growth in the past and in the coming decade. Hence it is not surprising that about 85% of Internet users surveyed claim using search engines and search services to find specific information. The same surveys show, however, that users are not satisfied with the performance of the current generation of search engines; the slow retrieval speed, communication delays, and poor quality of retrieved results (e.g., noise and broken links) are commonly cited problems. We discuss the development of new techniques targeted to resolve some of the problems associated with Web-based information retrieval,and speculate on future trends.

Categories and Subject Descriptors: G.1.3 [**Numerical Analysis**]: Numerical Linear Algebra—*Eigenvalues and eigenvectors* (direct and iterative methods); *Singular value decomposition*; *Sparse, structured and very large systems* (direct and iterative methods); G.1.1 [**Numerical Analysis**]: Interpolation; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; *Retrieval models*; *Search process*; H.m [**Information Systems**]: Miscellaneous

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Clustering, indexing, information retrieval, Internet, knowledge management, search engine, World Wide Web

1. INTRODUCTION

We review some notable studies on the growth of the Internet and on technologies useful for information search and retrieval on the Web. Writing about the Web is a challenging task for several reasons, of which we mention three. First, its dynamic nature guarantees that at least some portions of any manuscript on the subject will be out-of-date before it reaches the intended audience, particularly URLs that are referenced. Second, a comprehensive coverage of all of the important topics is impossible, because so many new ideas are constantly being proposed and are either quickly accepted into the Internet mainstream or rejected. Finally, as with any review paper, there is a strong bias

Authors' address: Tokyo Research Laboratory, IBM Research, 1623–14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242–8502, Japan; email: mei_kobayashi@jp.ibm.com; kohichi_takeda@jp.ibm.com.

CONTENTS

in presenting topics closely related to the authors' background, and giving only cursory treatment to those of which they are relatively ignorant. In an attempt to compensate for oversights and biases, references to relevant works that describe or review concepts in depth will be given whenever possible. This being said, we begin with references to several excellent books that cover a variety of topics in information management and retrieval. They include *Information Retrieval and Hypertext* [Agosti and Smeaton 1996]; *Modern Information Retrieval* [Baeza-Yates and Ribeiro-Neto 1999]; *Text Retrieval and Filtering: Analytic Models of Performance* [Losee 1998]; *Natural Language Information Retrieval* [Strzalkowski 1999]; and *Managing Gigabytes* [Witten et al. 1994]. Some older, classic texts, which are slightly outdated, include *Information Retrieval* [Frakes and Baeza-Yates 1992]; *Information Storage and Retrieval* [Korfhage 1997]; *Intelligent Multmedia Information Retrieval* [Maybury 1997]; *Introduction to Modern Information Retrieval* [Salton and McGill 1983]; and *Readings in Information Retrieval* [Jones and Willett 1977].

Additional references are to special journal issues on search engines on the Internet [Scientific American 1997]; digital libraries [CACM 1998]; digital libraries, representation and retrieval [IEEE 1996b]; the next generation graphical user interfaces (GUIs) [CACM 1994]; Internet technologies [CACM 1994; IEEE 1999]; and knowledge discovery [CACM 1999]. Some notable survey papers are those by Chakrabarti and Rajagopalan [1997]; Faloutsos and Oard [1995]; Feldman [1998]; Gudivada et al. [1997]; Leighton and Srivastava [1997]; Lawrence and Giles [1998b; 1999b]; and Raghavan [1997]. Extensive, up-to-date coverage of topics in Web-based information retrieval and knowledge management can be found in the proceedings of several conferences, such as: the *International World Wide Web Conferences* [WWW Conferences 2000] and the Association for Computing Machinery's Special Interest Group on Computer-Human Interaction [ACM SIGCHI] and Special Interest Group on Information Retrieval [ACM SIGIR] conferences <acm.org>. A list of papers and Web pages that review and compare Web search tools are maintained at several sites, including Boutell's World Wide Web FAQ <boutell.com/faq/>; Hamline University's <web.hamline.edu/administration/libraries/search/comparisons.html>; Kuhn's pages (in German) <gwdg.de/hkuhn1/pagesuch.html#vl>; Maire's pages (in French) <imaginet.fr/ime/search.htm>; Princeton University's <cs.princeton.edu/html/search.html>; U.C. Berkeley's <sunsite.berkeley.edu/help/searchdetails.html>; and Yahoo!'s pages on search engines <yahoo.com/computers and internet/internet/world wide web>. The historical development of information retrieval is documented in a number of sources: Baeza-Yates and Ribeiro-Neto [1999]; Cleverdon [1970]; Faloutsos and Oard [1995]; Salton [1970]; and van Rijsbergen [1979]. Historical accounts of the Web and Web search technologies are given in Berners-Lee et al. [1994] and Schatz [1997].

This paper is organized as follows. In the remainder of this section, we discuss and point to references on ratings of search engines and their features, the growth of information available on the Internet, and the growth in users. In the second section we present tools for Web-based information retrieval. These

include classical retrieval tools (which can be used as is or with enhancements specifically geared for Web-based applications), as well as a new generation of tools which have developed alongside the Internet. Challenges that must be overcome in developing and refining new and existing technologies for the Web environment are discussed. In the concluding section, we speculate on future directions in research related to Web-based information retrieval which may prove to be fruitful.

1.1 Ratings of Search Engines and their Features

About 85% of Web users surveyed claim to be using search engines or some kind of search tool to find specific information of interest. The list of publicly accessible search engines has grown enormously in the past few years (see, e.g., blueangels.net), and there are now lists of top-ranked query terms available online (see, e.g., <searchterms.com>). Since advertising revenue for search and portal sites is strongly linked to the volume of access by the public, increasing hits (i.e., demand for a site) is an extremely serious business issue. Undoubtedly, this financial incentive is serving as one the major impetuses for the tremendous amount of research on Web-based information retrieval.

One of the keys to becoming a popular and successful search engine lies in the development of new algorithms specifically designed for fast and accurate retrieval of valuable information. Other features that make a search or portal site highly competitive are unusually attractive interfaces, free email addresses, and free access time [Chandrasekaran 1998]. Quite often, these advantages last at most a few weeks, since competitors keep track of new developments (see, e.g., <portalhub.com> or <traffik.com>, which gives updates and comparisons on portals). And sometimes success can lead to unexpected consequences:

> *"Lycos, one of the biggest and most popular search engines, is legendary for its unavailability during work hours."* [Webster and Paul 1996]

There are many publicly available search engines, but users are not necessarily satisfied with the different formats for inputting queries, speeds of retrieval, presentation formats of the retrieval results, and quality of retrieved information [Lawrence and Giles 1998b]. In particular, speed (i.e., search engine search and retrieval time plus communication delays) has consistently been cited as "the most commonly experienced problem with the Web" in the biannual WWW surveys conducted at the Graphics, Visualization, and Usability Center of the Georgia Institute of Technology.[1] 63% to 66% of Web users in the past three surveys, over a period of a year-and-a-half were dissatisfied with the speed of retrieval and communication delay, and the problem appears to be growing worse. Even though 48% of the respondents in the April 1998 survey had upgraded modems in the past year, 53% of the respondents left a Web site while searching for product information because of "slow access." "Broken links" registered as the second most frequent problem in the same survey. Other studies also cite the number one and number two reasons for dissatisfaction as "*slow access*" and "*the inability to find relevant information*," respectively [Huberman and Lukose 1997; Huberman et al. 1998]. In this paper we elaborate on some of the causes of these problems and outline some promising new approaches being developed to resolve them.

It is important to remember that problems related to speed and access time may not be resolved by considering Web-based information access and retrieval as an isolated scientific problem. An August 1998 survey by Alexa Internet

[1]GVU's user survey (available at <gvu.gatech.edu/user surveys/>) is one of the more reliable sources on user data. Its reports have been endorsed by the World Wide Web Consortium (W3C) and INRIA.

<alexa.com/company/inthenews/webfacts.html> indicates that 90% of all Web traffic is spread over 100,000 different hosts, with 50% of all Web traffic headed towards the top 900 most popular sites. Effective means of managing uneven concentration of information packets on the Internet will be needed in addition to the development of fast access and retrieval algorithms.

The volume of information on search engines has exploded in the past year. Some valuable resources are cited below. The University of California at Berkeley has extensive Web pages on "how to choose the search tools you need" <lib.berkeley.edu/teachinglib/guides/internet/toolstables.html>. In addition to general advice on conducting searches on the Internet, the pages compare features such as size, case sensitivity, ability to search for phrases and proper names, use of Boolean logic terms, ability to require or exclude specified terms, inclusion of multilingual features, inclusion of special feature buttons (e.g., *"more like this," "top 10 most frequently visited sites on the subject,"* and *"refine"*) and exclusion of pages updated prior to a user-specified date of several popular search engines such as those of Alta Vista <altavista.com>; HotBot <hotbot.com>; Lycos Pro Power Search <lycos.com>; Excite <excite.com>; Yahoo! <yahoo.com>; Infoseek <infoseek.com>; Disinformation <disinfo.com>; and Northern Light <nlsearch.com>.

The work of Lidsky and Kwon [1997] is an opinionated but informative resource on search engines. It describes 36 different search engines and rates them on specific details of their search capabilities. For instance, in one study, searches are divided into five categories: (1) simple searches; (2) custom searches; (3) directory searches; (4) current news searches; and (5) Web content. The five categories of search are evaluated in terms of power and ease of use. Variations in ratings sometimes differ substantially for a given search engine. Similarly, query tests are conducted according to five criteria: (1) simple queries; (2) customized queries; (3) news queries; (4) duplicate elimination; and (5) dead link elimination. Once again, variations in the ratings sometimes differ substantially for a given search engine. In addition to ratings, the authors give charts on search indexes and directories associated with twelve of the search engines, and rate them in terms of specific features for complex searches and content. The data indicate that as the number of people using the Internet and Web has grown, user types have diversified and search engine providers have begun to target more specific types of users and queries with specialized and tailored search tools.

Web Search Engine Watch <searchenginewatch.com/webmasters/features.html> posts extensive data and ratings of popular search engines according to features such as size, pages crawled per day, freshness, and depth. Some other useful online sources are home pages on search engines by the Gray <mit.people.edu/mkgray/net>; Information Today <infotoday.com/searcher/jun/story2.htm>; Kansas City Public Library <kcpl.lib.mo.us/search/srchengines.htm>; Koch <ub2.lu.se/desire/radar/lit-about-search-services.html>; Northwestern University Library <library.nwu.edu/resources/internet/search/evaluate.html>; and Notes of Search Engine Showdown <imtnet/notes/search/index.html>. Data on international use of the Web and Internet is posted at the NUA Internet Survey home page <nua.ie/surveys>.

A note of caution: in digesting the data in the paragraphs above and below, published data on the Internet and the Web are very difficult to measure and verify. GVU offers a solid piece of advice on the matter:

> *"We suggest that those interested in these (i.e., Internet/WWW statistics and demographics) statistics should consult several sources; these numbers can be difficult to measure and results may vary between different sources."* [GVU's WWW user survey]

Although details of data from different

popular sources vary, overall trends are fairly consistently documented. We present some survey results from some of these sources below.

1.2 Growth of the Internet and the Web

Schatz [1997] of the National Center for Supercomputing Applications (NCSA) estimates that the number of Internet users increased from 1 million to 25 million in the five years leading up to January of 1997. Strategy Alley [1998] gives a number of statistics on Internet users: Matrix Information and Directory Services (MIDS), an Internet measurement organization, estimated there were 57 million users on the consumer Internet worldwide in April of 1998, and that the number would increase to 377 million by 2000; Morgan Stanley gives the estimate of 150 million in 2000; and Killen and Associates give the estimate as 250 million in 2000. Nua's surveys <nua.ie/surveys> estimates the figure as 201 million worldwide in September of 1999, and more specifically by region: 1.72 million in Africa; 33.61 in the Asia/Pacific region; 47.15 in Europe; 0.88 in the Middle East; 112.4 in Canada and the U.S.; and 5.29 in Latin America. Most data and projections support continued tremendous growth (mostly exponential) in Internet users, although precise numerical values differ.

Most data on the amount of information on the Internet (i.e., volume, number of publicly accessible Web pages and hosts) show tremendous growth, and the sizes and numbers appear to be growing at an exponential rate. Lynch has documented the explosive growth of Internet hosts; the number of hosts has been roughly doubling every year. For example, he estimates that it was 1.3 million in January of 1993, 2.2 million in January of 1994, 4.9 million in January of 1995, and 9.5 million in January of 1996. His last set of data is 12.9 million in July of 1996 [Lynch 1997]. Strategy Alley [1998] cites similar figures: "*Since 1982, the number of hosts has doubled every year*." And an article by the editors of the *IEEE Internet Computing Magazine* states that exponential growth of Internet hosts was observed in separate studies by several experts [IEEE 1998a], such as Mark Lottor of Network Wizards <nw.com>; Mirjan Kühne of the RIPE Network Control Center <.ripe.net> for a period of over ten years; Samarada Weerahandi of Bellcore on his home page on Internet hosts <ripe.net> for a period of over five years in Europe; and John Quarterman of Matrix Information and Directory Services <mids.org>.

The number of publicly accessible pages is also growing at an aggressive pace. Smith [1973] estimates that in January of 1997 there were 80 million public Web pages, and that the number would subsequently double annually. Bharat and Broder [1998] estimated that in November of 1997 the total number of Web pages was over 200 million. If both of these estimates for number of Web pages are correct, then the rate of increase is higher than Smith's prediction, i.e., it would be more than double per year. In a separate estimate [Monier 1998], the chief technical officer of AltaVista estimated that the volume of publicly accessible information on the Web has grown from 50 million pages on 100,000 sites in 1995 to 100 to 150 million pages on 600,000 sites in June of 1997. Lawrence and Giles summarize Web statistics published by others: 80 million pages in January of 1997 by the Internet Archive [Cunningham 1997], 75 million pages in September of 1997 by Forrester Research Inc. [Guglielmo 1997], Monier's estimate (mentioned above), and 175 million pages in December 1997 by Wired Digital. Then they conducted their own experiments to estimate the size of the Web and concluded that:

> *"it appears that existing estimates significantly underestimate the size of the Web."* [Lawrence and Giles 1998b]

Follow-up studies by Lawrence and Giles [1999a] estimate that the number of publicly indexable pages on the Web

at that time was about 800 million pages (with a total of 6 terabytes of text data) on about 3 million servers (Lawrence's home page: <neci.nec.cim/lawrence/papers.html>). On Aug. 31 1998, Alexa Internet announced its estimate of 3 terabytes or 3 million megabytes for the amount of information on the Web, with 20 million Web content areas; a content area is defined as top-level pages of sites, individual home pages, and significant subsections of corporate Web sites. Furthermore, they estimate a doubling of volume every eight months.

Given the enormous volume of Web pages in existence, it comes as no surprise that Internet users are increasingly using search engines and search services to find specific information. According to Brin and Paige, the World WideWebWorm (homepages: <cs.colorado.edu/wwww> and <guano.cs.colorado.edu/wwww>) claims to have handled an average of 1,500 queries a day in April 1994, and AltaVista claims to have handled 20 million queries in November 1997. They believe that

> *"it is likely that top search engines will handle hundreds of millions (of queries) per day by the year 2000."* [Brin and Page 1998]

The results of GVU's April 1998 WWW user survey indicate that about 86% of people now find a useful Web site through search engines, and 85% find them through hyperlinks in other Web pages; people now use search engines as much as surfing the Web to find information.

1.3 Evaluation of Search Engines

Several different measures have been proposed to quantitatively measure the performance of classical information retrieval systems (see, e.g., Losee [1998]; Manning and Schutze [1999]), most of which can be straightforwardly extended to evaluate Web search engines. However, Web users may have a tendency to favor some performance issues more strongly than traditional users of information retrieval systems. For example, interactive *response times* appear to be at the top of the list of important issues for Web users (see Section 1.1) as well as the number of valuable sites listed in the first page of retrieved results (i.e., ranked in the top 8, 10, or 12), so that the *scroll down* or *next page* button do not have to be invoked to view the most valuable results.

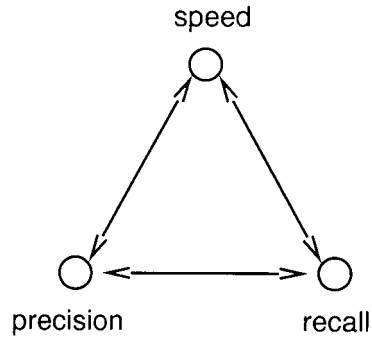

Figure 1. Three way trade-off in search engine performance: (1) speed of retrieval, (2) precision, and (3) recall.

Some traditional measures of information retrieval system performance are recognized in modified form by Web users. For example, a basic model from traditional retrieval systems recognizes a three way trade-off between the speed of information retrieval, precision, and recall (which is illustrated in Figure 1). This trade-off becomes increasingly difficult to balance as the number of documents and users of a database escalate. In the context of information retrieval, *precision* is defined as the ratio of relevant documents to the number of retrieved documents:

$$\text{precision} = \frac{\text{number of relevant documents}}{\text{number of retrieved documents}},$$

and *recall* is defined as the proportion of relevant documents that are retrieved:

$$\text{recall} = \frac{\text{number of relevant, retrieved documents}}{\text{total number of relevant documents}}.$$

Most Web users who utilize search engines are not so much interested in the traditional measure of precision as the precision of the results displayed in the first page of the list of retrieved documents, before a "*scroll*" or "*next page*" command is used. Since there is little hope of actually measuring the recall rate for each Web search engine query and retrieval job—and in many cases there may be too many relevant pages—a Web user would tend to be more concerned about retrieving and being able to identify only very highly valuable pages. Kleinberg [1998] recognizes the importance of finding the most information rich, or *authority* pages. *Hub* pages, i.e., pages that have links to many *authority* pages are also recognized as being very valuable. A Web user might substitute recall with a modified version in which the recall is computed with respect to the set of hub and authority pages retrieved in the top 10 or 20 ranked documents (rather than all related pages). Details of an algorithm for retrieving authorities and hubs by Kleinberg [1998] is given in Section 2.4 of this paper.

Hearst [1999] notes that the user interface, i.e., the quality of human-computer interaction, should be taken into account when evaluating an information retrieval system. Nielsen [1993] advocates the use of qualitative (rather than quantitative) measures to evaluate information retrieval systems. In particular, user satisfaction with the system interface as well as satisfaction with retrieved results as a whole (rather than statistical measures) is suggested. Westera [1996] suggests some query formats for benchmarking search engines, such as: single keyword search; plural search capability; phrase search; Boolean search (with proper noun); and complex Boolean. In the next section we discuss some of the differences and similarities in classical and Internet-based search, access and retrieval of information.

Hawking et al. [1999] discusses evaluation studies of six text retrieval conferences (TREC) U.S. National Institute of Standards and Technology (NIST) search engines <trec.nist.gov>. In particular, they examine answers to questions such as *"Can link information result in better rankings?"* and *"Do longer queries result in better answers?"*

2. TOOLS FOR WEB-BASED RETRIEVAL AND RANKING

Classical retrieval and ranking algorithms developed for isolated (and sometimes static) databases are not necessarily suitable for Internet applications. Two of the major differences between classical and Web-based retrieval and ranking problems and challenges in developing solutions are the number of simultaneous users of popular search engines and the number of documents that can be accessed and ranked. More specifically, the number of simultaneous users of a search engine at a given moment cannot be predicted beforehand and may overload a system. And the number of publicly accessible documents on the Internet exceeds those numbers associated with classical databases by several orders of magnitude. Furthermore, the number of Internet search engine providers, Web users, and Web pages is growing at a tremendous pace, with each average page occupying more memory space and containing different types of multimedia information such as images, graphics, audio, and video.

There are other properties besides the number of users and size that set classical and Web-based retrieval problems apart. If we consider the set of all Web pages as a gigantic database, this set is very different from a classical database with elements that can be organized, stored, and indexed in a manner that facilitates fast and accurate retrieval using a well-defined format for input queries. In Web-based retrieval, determining which pages are valuable enough to index, weight, or cluster and carrying out the tasks efficiently, while maintaining a reasonable degree of

accuracy considering the ephemeral nature of the Web, is an enormous challenge. Further complicating the problem is the set of appropriate input queries; the best format for inputting the queries is not fixed or known. In this section we examine indexing, clustering, and ranking algorithms for documents available on the Web and user interfaces for protoype IR systems for the Web.

2.1 Indexing

The American Heritage Dictionary (1976) defines *index* as follows:

> **(in · dex) 1.** Anything that serves to guide, point out or otherwise facilitate reference, as: **a.** An alphabetized listing of names, places, and subjects included in a printed work that gives for each item the page on which it may be found. **b.** A series of notches cut into the edge of a book for easy access to chapters or other divisions. **c.** Any table, file, or catalogue.

Although the term is used in the same spirit in the context of retrieval and ranking, it has a specific meaning. Some definitions proposed by experts are "The most important of the tools for information retrieval is the *index*—a collection of terms with pointers to places where information about documents can be found" [Manber 1999]; "*indexing* is building a data structure that will allow quick seaching of the text" [Baeza-Yates 1999]; or "*the act of assigning index terms to documents, which are the objects to be retrieved*" [Korfhage 1997]; "An *index term* is a (document) word whose semantics helps in remembering the document's main themes" [Baeza-Yates and Ribeiro-Neto 1999]. Four approaches to indexing documents on the Web are (1) human or manual indexing; (2) automatic indexing; (3) intelligent or agent-based indexing; and (4) metadata, RDF, and annotation-based indexing. The first two appear in many classical texts, while the latter two are relatively new and promising areas of study. We first give an overview of Web-based indexing, then describe or give references to the various approaches.

Indexing Web pages to facilitate retrieval is a much more complex and challenging problem than the corresponding one associated with classical databases. The enormous number of existing Web pages and their rapid increase and frequent updating makes straightforward indexing, whether by human or computer-assisted means, a seemingly impossible, Sisyphean task. Indeed, most experts agree that, at a given moment, a significant portion of the Web is not recorded by the indexer of any search engine. Lawrence and Giles estimated that, in April 1997, the lower bound on indexable Web pages was 320 million, and a given individual search engine will have indexed between 3% to 34% of the possible total [Lawrence and Giles 1998b]. They also estimated that the extent of overlap among the top six search engines is small and their collective coverage was only around 60%; the six search engines are HotBot, AltaVista, Northern Light, Excite, Infoseek, and Lycos. A follow up study for the period February 2–28, 1999, involving the top 11 search engines (the six above plus Snap <snap.com>; Microsoft <msn.com>; Google <google.com>; Yahoo!; and Euroseek <euroseek.com>) indicates that we are losing the indexing race. A far smaller proportion of the Web is now indexed with no engine covering more than 16% of the Web. Indexing appears to have become more important than ever, since 83% of sites contained commercial content and 6% contained scientific or educational content [Lawrence and Giles 1999a].

Bharat and Broder estimated in November 1997 that the number of pages indexed by HotBot, AltaVista, Excite, and Infoseek were 77 million, 100 million, 32 million, and 17 million, respectively. Furthermore, they believe that the union of these pages is around 160 million pages, i.e., about 80% of the 200 million total accessible pages they believe

existed at that time. Their studies indicate that there is little overlap in the indexing coverage, more specifically, less than 1.4% (i.e., 2.2 million) of the 160 million indexed pages were covered by all four of the search engines. Melee's Indexing Coverage Analysis (MICA) Reports <melee.com/mica/index.html> provides a weekly update on indexing coverage and quality by a few, select, search engines that claim to index "at least one fifth of the Web." Other studies on estimating the extent of Web pages that have been indexed by popular search engines include Baldonado and Winograd [1997]; Hernandez [1996]; Hernandez and Stolfo [1995]; Hylton [1996]; Monge and Elkan [1998]; Selberg and Etzioni [1995a]; and Silberschatz et al. [1995].

In addition to the sheer volume of documents to be processed, indexers must take into account other complex issues, for example, Web pages are not constructed in a fixed format; the textual data is riddled with an unusually high percentage of typos—the contents usually contain nontextual multimedia data, and updates to the pages are made at different rates. For instance, preliminary studies documented in Navarro [1998] indicate that on the average site 1 in 200 common words and 1 in 3 foreign surnames are misspelled. Brake [1997] estimates that the average page of text remains unchanged on the Web for about 75 days, and Kahle estimates that 40% of the Web changes every month. Multiple copies of identical or near-identical pages are abundant; for example, FAQs [2] postings, mirror sites, old and updated versions of news, and newspaper sites. Broder et al. [1997] and Shivakumar and García-Molina [1998] estimate that 30% of Web pages are duplicates or near-duplicates. Tools for removing redundant URLs or URLs of near and perfectly identical sites have been investigated by Baldonado and Winograd [1997]; Hernandez [1996]; Hernandez and Stolfo [1995]; Hylton [1996]; Monge and Elkan [1998]; Selberg and Etzioni [1995a]; and Silberschatz et al. [1995].

Henzinger et al. [1999] suggested a method for evaluating the quality of pages in a search engine's index. In the past, the volume of pages indexed was used as the primary measurement of Web page indexers. Henzinger et al. suggest that the quality of the pages in a search engine's index should also be considered, especially since it has become clear that no search engine can index all documents on the Web, and there is very little overlap between the indexed pages of major search engines. The idea of Henzinger's method is to evaluate the quality of Web pages according to its *indegree* (an evaluation measure based on how many other pages point to the Web page under consideration [Carriere and Kazman 1997]) and *PageRank* (an evaluation measure based on how many other pages point to the Web page under consideration, as well as the value of the pages pointing to it [Brin and Page 1998; Cho et al. 1998]).

The development of effective indexing tools to aid in filtering is another major class of problems associated with Web-based search and retrieval. Removal of spurious information is a particularly challenging problem, since a popular information site (e.g., newsgroup discussions, FAQ postings) will have little value to users with no interest in the topic. Filtering to block pornographic materials from children or for censorship of culturally offensive materials is another important area for research and business devlopment. One of the promising new approaches is the use of *metadata*, i.e., summaries of Web page content or sites placed in the page for aiding automatic indexers.

[2]FAQs, or frequently asked questions, are essays on topics on a wide range of interests, with pointers and references. For an extensive list of FAQs, see <cis.ohio-state.edu/hypertext/faq/usenet/faq-list.html> and <faq.org>.

2.1.1 *Classical Methods*. Manual indexing is currently used by several commercial, Web-based search engines, e.g., Galaxy <galaxy.einet.net>; GNN: Whole Internet Catalog <e1c.gnn.com/gnn/wic/index.html>; Infomine <lib-www.ucr.edu>; KidsClick! <sunsite.berkeley.edu/kidsclick!/>; LookSmart <looksmart.com>; Subject Tree <bubl.bath.ac.uk/bubl/cattree.html>; Web Developer's Virtual Library <stars.com>; World-Wide Web Virtual Library Series Subject Catalog <w3.org/hypertext/datasources/bysubject/overview.html>; and Yahoo!. The practice is unlikely to continue to be as successful over the next few years, since, as the volume of information available over the Internet increases at an ever greater pace, manual indexing is likely to become obsolete over the long term. Another major drawback with manual indexing is the lack of consistency among different professional indexers; as few as 20% of the terms to be indexed may be handled in the same manner by different individuals [Korfhage 1997, p. 107], and there is noticeable inconsistency, even by a given individual [Borko 1979; Cooper 1969; Jacoby and Slamecka 1962; Macskassys et al. 1998; Preschel 1972; and Salton 1969].

Though not perfect, compared to most automatic indexers, human indexing is currently the most accurate because experts on popular subjects organize and compile the directories and indexes in a way which (they believe) facilitates the search process. Notable references on conventional indexing methods, including automatic indexers, are Part IV of Soergel [1985]; Jones and Willett [1977]; van Rijsbergen [1977]; and Witten et al. [1994, Chap. 3]. Technological advances are expected to narrow the gap in indexing quality between human and machine-generated indexes. In the future, human indexing will only be applied to relatively small and static (or near static) or highly specialized data bases, e.g., internal corporate Web pages.

2.1.2 *Crawlers/Robots*. Scientists have recently been investigating the use of *intelligent agents* for performing specific tasks, such as indexing on the Web [AI Magazine 1997; Baeza-Yates and Ribeiro-Neto 1999]. There is some ambiguity concerning proper terminology to describe these agents. They are most commonly referred to as crawlers, but are also known as ants, automatic indexers, bots, spiders, Web robots (Web robot FAQ <info.webcrawler.com/mak/projects/robots/faq.html>), and worms. It appears that some of the terms were proposed by the inventors of a specific tool, and their subsequent use spread to more general applications of the same genre.

Many search engines rely on automatically generated indices, either by themselves or in combination with other technologies, e.g., Aliweb <nexor.co.uk/public/aliweb/aliweb.html>; AltaVista; Excite; Harvest <harvest.transarc.com>; HotBot; Infoseek; Lycos; Magellan <magellan.com>; MerzScope <merzcom.com>; Northern Light; Smart Spider <engsoftware.com>; Webcrawler <webcrawler.com/>; and World Wide Web Worm. Although most of Yahoo!'s entries are indexed by humans or acquired through submissions, it uses a robot to a limited extent to look for new announcements. Examples of highly specialized crawlers include Argos <argos.evansville.edu> for Web sites on the ancient and medieval worlds; CACTVS Chemistry Spider <schiele.organik.uni-erlangen.de/cactvs/spider.html> for chemical databases; MathSearch <maths.usyd.edu.au:8000/mathsearch.html> for English mathematics and statistics documents; NEC-MeshExplorer <netplaza.biglobe.or.jp/keyword.html> for the NETPLAZA search service owned by the NEC Corporation; and Social Science Information Gateway (SOSIG) <scout.cs.wisc.edu/scout/mirrors/sosig> for resources in the social sciences. Crawlers that index documents in limited environments include LookSmart <looksmart.com/> for a 300,000 site database of rated and reviewed sites; Robbie

the Robot, funded by DARPA for education and training purposes; and UCSD Crawl <www.mib.org/ ucsdcrawl> for UCSD pages. More extensive lists of intelligent agents are available on The Web Robots Page <info.webcrawler.com/mak/projects/robots/active/html/type.html>; and on Washington State University's robot pages <wsulibs.wsu.edu/general/robots.htm>.

To date, there are three major problems associated with the use of robots: (1) some people fear that these agents are too invasive; (2) robots can overload system servers and cause systems to be virtually frozen; and (3) some sites are updated at least several times per day, e.g., approximately every 20 minutes by CNN <cnn.com> and Bloomberg <bloomberg.com>, and every few hours by many newspaper sites [Carl 1995] (article home page <info.webcrawler.com/mak/projects/robots/threat-or-treat.html>); [Koster 1995]. Some Web sites deliberately keep out spiders; for example, the *New York Times* <nytimes.com> requires users to pay and fill out a registration form; CNN used to exclude search spiders to prevent distortion of data on the number of users who visit the site; and the online catalogue of the British Library <portico.bl.uk> only allows access to users who have filled out an online query form [Brake 1997]. System managers of these sites must keep up with the new spider and robot technologies in order to develop their own tools to protect their sites from new types of agents that intentionally or unintentionally could cause mayhem.

As a working compromise, Kostner has proposed a *robots exclusion standard* ("A standard for robots exclusion," ver.1:<info.webcrawler.com/mak/projects/robots/exclusion.html>; ver. 2: <info.webcrawler.com/mak/projects/robots/norobot.html>), which advocates blocking certain types of searches to relieve overload problems. He has also proposed guidelines for *robot design* ("Guidelines for robot writers" (1993) <info.webcrawler.com/mak/projects/robots/guidelines.html>). It is important to note that robots are not always the root cause of network overload; sometimes human user overload causes problems, which is what happened at the CNN site just after the announcement of the O.J. Simpson trial verdict [Carl 1995]. Use of the exclusion standard is strictly voluntary, so that Web masters have no guarantee that robots will not be able to enter computer systems and create havoc. Arguments in support of the exclusion standard and discussion on its effectiveness are given in Carl [1995] and Koster [1996].

2.1.3 *Metadata, RDF, and Annotations.*

> *"What is metadata? The Macquarie dictionary defines the prefix 'meta-' as meaning 'among,' 'together with,' 'after' or 'behind.' That suggests the idea of a 'fellow traveller ': that metadata is not fully fledged data, but it is a kind of fellow-traveller with data, supporting it from the sidelines. My definition is that 'an element of metadata describes an information resource or helps provide access to an information resource.'"* [Cathro 1997]

In the context of Web pages on the Internet, the term "*metadata*" usually refers to an invisible file attached to a Web page that facilitates collection of information by automatic indexers; the file is invisible in the sense that it has no effect on the visual appearance of the page when viewed using a standard Web browser.

The World Wide Web (W3) Consortium <w3.org> has compiled a list of resources on information and standardization proposals for metadata (W3 metadata page <w3.org/metadata>. A number of metadata standards have been proposed for Web pages. Among them, two well-publicized, solid efforts are the Dublin Core Metadata standard: home page <purl.oclc.org/metadata/dublin core> and the Warwick framework: article home page <dlib.org/dlib/july96/lagoze/07lagoze.html> [Lagoze 1996]. The Dublin Core is a 15-element metadata element set proposed to facilitate fast and accurate information retrieval on the Internet. The elements are title; creator; subject; description;

publisher; contributors; date; resource type; format; resource identifier; source; language; relation; coverage; and rights. The group has also developed methods for incorporating the metadata into a Web page file. Other resources on metadata include Chapter 6 of Baeza-Yates and Ribeiro-Neto [1999] and Marchionini [1999]. If the general public adopts and increases use of a simple metadata standard (such as the Dublin Core), the precision of information retrieved by search engines is expected to improve substantially. However, widespread adoption of a standard by international users is dubious.

One of the major drawbacks of the simplest type of metadata for labeling HTML documents, called *metatags*, is they can only be used to describe contents of the document to which they are attached, so that managing collections of documents (e.g., directories or those on similar topics) may be tedious when updates to the entire collection are made. Since a single command cannot be used to update the entire collection at once, documents must be updated one-by-one. Another problem is when documents from two or more different collections are merged to form a new collection. When two or more collections are merged, inconsistent use of metatags may lead to confusion, since a metatag might be used in different collections with entirely different meanings. To resolve these issues, the W3 Consortium proposed in May 1999 that the Resource Description Framework (RDF) be used as the metadata coding scheme for Web documents (W3 Consortium RDF homepage <w3.org/rdf>. An interesting associated development is IBM's XCentral <ibm.com/developer/xml>, the first search engine that indexes XML and RDF elements.

Metadata places the responsibility of aiding indexers on the Web page author, which is reasonable if the author is a responsible person wishing to advertise the presence of a page to increase legitimate traffic to a site. Unfortunately, not all Web page authors are fair players. Many unfair players maintain sites that can increase advertising revenue if the number of visitors is very high or charging a fee per visit for access to pornographic, violent, and culturally offensive materials. These sites can attract a large volume of visitors by attaching metadata with many popular keywords. Development of reliable filtering services for parents concerned about their children's surfing venues is a serious and challenging problem.

Spamming, i.e., excessive, repeated use of key words or "hidden" text purposely inserted into a Web page to promote retrieval by search engines, is related to, but separate from, the unethical or deceptive use of metadata. Spamming is a new phenomenon that appeared with the introduction of search engines, automatic indexers, and filters on the Web [Flynn 1996; Liberatore 1997]. Its primary intent is to outsmart these automated software systems for a variety of purposes; spamming has been used as an advertising tool by entrepreneurs, cult recruiters, egocentric Web page authors wanting attention, and technically well-versed, but unbalanced, individuals who have the same sort of warped mentality as inventors of computer viruses. A famous example of hidden text spamming is the embedding of words in a black background by the Heaven's Gate cult. Although the cult no longer exists, its home page is archived at the sunspot.net site <sunspot.net/news/special/heavensgatesite>, a technique known as *font color spamming* [Liberatore 1997]. We note that the term *spamming* has a broader meaning, related to receiving an excessive amount of email or information. An excellent, broad overview of the subject is given in Cranor and LaMacchia [1998]. In our context, the specialized terms *spam-indexing*, *spam-dexing*, or *keyword spamming* are more precise.

Another tool related to metadata is *annotation*. Unlike metadata, which is created and attached to Web documents by the author for the specific purpose of

aiding indexing, annotations include a much broader class of data to be attached to a Web document [Nagao and Hasida 1998; Nagao et al. 1999]. Three examples of the most common annotations are linguistic annotation, commentary (created by persons other than the author), and multimedia annotation. Linguistic annotation is being used for automatic summarization and content-based retrieval. Commentary annotation is used to annotate nontextual multimedia data, such as image and sound data plus some supplementary information. Multimedia annotation generally refers to text data, which describes the contents of video data (which may be downloadable from the Web page). An interesting example of annotation is the attachment of comments on Web documents by people other than the document author. In addition to aiding indexing and retrieval, this kind of annotation may be helpful for evaluating documents.

Despite the promise that metadata and annotation could facilitate fast and accurate search and retrieval, a recent study for the period February 2–28, 1999 indicates that metatags are only used on 34% of homepages, and only 0.3% of sites use the Dublin Core metadata standard [Lawrence and Giles 1999a]. Unless a new trend towards the use of metadata and annotations develops, its usefulness in information retrieval may be limited to very large, closed data owned by large corporations, public institutions, and governments that choose to use it.

2.2 Clustering

Grouping similar documents together to expedite information retrieval is known as *clustering* [Anick and Vaithyanathan 1997; Rasmussen 1992; Sneath and Sokal 1973; Willett 1988]. During the information retrieval and ranking process, two classes of similarity measures must be considered: the similarity of a document and a query and the similarity of two documents in a database. The similarity of two documents is important for identifying groups of documents in a database that can be retrieved and processed together for a given type of user input query.

Several important points should be considered in the development and implementation of algorithms for clustering documents in very large databases. These include identifying relevant attributes of documents and determining appropriate weights for each attribute; selecting an appropriate clustering method and similarity measure; estimating limitations on computational and memory resources; evaluating the reliability and speed of the retrieved results; facilitating changes or updates in the database, taking into account the rate and extent of the changes; and selecting an appropriate search algorithm for retrieval and ranking. This final point is of particularly great concern for Web-based searches.

There are two main categories of clustering: *hierarchical* and *nonhierarchical*. Hierarchical methods show greater promise for enhancing Internet search and retrieval systems. Although details of clustering algorithms used by major search engines are not publicly available, some general approaches are known. For instance, Digital Equipment Corporation's Web search engine *AltaVista* is based on clustering. Anick and Vaithyanathan [1997] explore how to combine results from latent semantic indexing (see Section 2.4) and analysis of phrases for context-based information retrieval on the Web.

Zamir et al. [1997] developed three clustering methods for Web documents. In the *word-intersection clustering* method, words that are shared by documents are used to produce clusters. The method runs in $O(n^2)$ time and produces good results for Web documents. A second method, *phrase-intersection clustering*, runs in $O(n \log n)$ time is at least two orders of magnitude faster than methods that produce comparable clusters. A third method, called *suffix tree clustering* is detailed in Zamir and Etzioni [1998].

Modha and Spangler [2000] developed a clustering method for hypertext documents, which uses *words* contained in the document, *outlinks* from the document, and *in-links* to the document. Clustering is based on six information nuggets, which the authors dubbed *summary, breakthrough, review, keywords, citation*, and *reference*. The first two are derived from the words in the document, the next two from the outlinks, and the last two from the in-links.

Several new approaches to clustering documents in data mining applications have recently been developed. Since these methods were specifically designed for processing very large data sets, they may be applicable with some modifications to Web-based information retrieval systems. Examples of some of these techniques are given in Agrawal et al. [1998]; Dhillon and Modha [1999; 2000]; Ester et al. [1995a; 1995b; 1995c]; Fisher [1995]; Guha et al. [1998]; Ng and Han [1994]; and Zhang et al. [1996]. For very large databases, appropriate parallel algorithms can speed up computation [Omiecinski and Scheuermann 1990].

Finally, we note that clustering is just one of several ways of organizing documents to facilitiate retrieval from large databases. Some alternative methods are discussed in Frakes and Baeza-Yates [1992]. Specific examples of some methods designed specifically for facilitating Web-based information retrieval are evaluation of significance, reliability, and topics covered in a set of Web pages based on analysis of the hyperlink structures connecting the pages (see Section 2.4); and identification of cyber communities with expertise in subject(s) based on user access frequency and surfing patterns.

2.3 User Interfaces

Currently, most Web search engines are text-based. They display results from input queries as long lists of pointers, sometimes with and sometimes without summaries of retrieved pages. Future commercial systems are likely to take advantage of small, powerful computers, and will probably have a variety of mechanisms for querying nontextual data (e.g., hand-drawn sketches, textures and colors, and speech) and better user interfaces to enable users to visually manipulate retrieved information [Card et al. 1999; Hearst 1997; Maybury and Wahlster 1998; Rao et al. 1993; Tufte 1983]. Hearst [1999] surveys visualization interfaces for information retrieval systems, with particular emphasis on Web-based systems. A sampling of some exploratory works being conducted in this area are described below. These interfaces and their display systems, which are known under several different names (e.g., dynamic querying, information outlining, visual information seeking), are being developed at universities, government, and private research labs, and small venture companies worldwide.

2.3.1 *Metasearch Navigators.* A very simple tool developed to exploit the best features of many search engines is the metasearch navigator. These navigators allow simultaneous search of a set of other navigators. Two of the most extensive are Search.com <search.com/>, which can utilize the power of over 250 search engines, and INFOMINE <lib-www.ucr.edu/enbinfo.html>, which utilizes over 90. Advanced metasearch navigators have a single input interface that sends queries to all (or only user selected search engines), eliminates duplicates, and then combines and ranks returned results from the different search engines. Some fairly simple examples available on the Web are 2ask <web.gazeta.pl/miki/search/2ask-anim.html>; ALL-IN-ONE <albany.net/allinone/>; EZ-Find at The River <theriver.com/theRiver/explore/ezfind.html>; IBM InfoMarket Service <infomkt.ibm.com/>; Inference Find <inference.com/infind/>; Internet Sleuth <intbc.com/sleuth>; MetaCrawler <metacrawler.cs.washington.edu:8080/>; and SavvySearch <cs.colostat.edu/dreiling/smartform.html> and <guaraldi.cs.colostate.edu:2000/> [Howe and Dreilinger 1997].

2.3.2 *Web-Based Information Outlining/Visualization.* Visualization tools specifically designed to help users understand websites (e.g., their directory structures, types of information available) are being developed by many private and public research centers [Nielsen 1997]. Overviews of some of these tools are given in Ahlberg and Shneiderman [1994]; Beaudoin et al. [1996]; Bederson and Hollan [1994]; Gloor and Dynes [1998]; Lamping et al. [1995]; Liechti et al. [1998]; Maarek et al. [1997]; Munzner and Burchard [1995]; Robertson et al. [1991]; and Tetranet Software Inc. [1998] <tetranetsoftware.com>. Below we present some examples of interfaces designed to facilitate general information retrieval systems, we then present some that were specifically designed to aid retrieval on the Web.

Shneiderman [1994] introduced the term *dynamic queries* to describe interactive user control of visual query parameters that generate a rapid, updated, animated visual display of database search results. Some applications of the dynamic query concept are systems that allow real estate brokers and their clients to locate homes based on price, number of bedrooms, distance from work, etc. [Williamson and Shneiderman 1992]; locate geographical regions with cancer rates above the national average [Plaisant 1994]; allow dynamic querying of a chemistry table [Ahlberg and Shneiderman 1997]; with an interface to enable users to explore UNIX directories through dynamic queries [Liao et al. 1992]: Visual presentation of query components; visual presentation of results; rapid, incremental, and reversible actions; selection by pointing (not typing); and immediate and continuous feedback are features of the systems. Most graphics hardware systems in the mid-1990's were still too weak to provide adequate real-time interaction, but faster algorithms and advances in hardware should increase system speed up in the future.

Williams [1984] developed a user interface for information retrieval systems to "aid users in formulating a query." The system, *RABBIT III*, supports interactive refinement of queries by allowing users to critique retrieved results with labels such as "*require*" and "*prohibit*." Williams claims that this system is particularly helpful to naïve users "with only a vague idea of what they want and therefore need to be guided in the formulation/reformulation of their queries . . . (or) who have limited knowledge of a given database or who must deal with a multitude of databases."

Hearst [1995] and Hearst and Pederson [1996] developed a visualization system for displaying information about a document and its contents, e.g., its length, frequency of term sets, and distribution of term sets within the document and to each other. The system, called TileBars, displays information about a document in the form of a two-dimensional rectangular bar with even-sized tiles lying next to each other in an orderly fashion. Each tile represents some feature of the document; the information is encoded as a number whose magnitude is represented in grayscale.

Cutting et al. [1993] developed a system called Scatter/Gather to allow users to cluster documents interactively, browse the results, select a subset of the clusters, and cluster this subset of documents. This process allows users to iteratively refine their search. BEAD [Chalmers and Chitson 1992]; Galaxy of News [Rennison 1994]; and ThemeScapes [Wise et al. 1995] are some of the other systems that show graphical displays of clustering results.

Baldonado [1997] and Baldonado and Winograd [1997] developed an interface for exploring information on the Web across heterogeneous sources, e.g., search services such as Alta Vista, bibliographic search services such as Dialog, a map search service and a video search service. The system, called SenseMaker, can "bundle" (i.e., cluster) similar types of retrieved data according to user specified "bundling criteria" (the

criteria must be selected from a fixed menu provided by SenseMaker). Examples of available bundling criteria for a URL type include "(1) bundling results whose URLs refer to the same site; (2) bundling results whose URLs refer to the same collection at a site; and (3) not bundling at all." The system allows users to select from several criteria to view retrieved results, e.g., according to the URL, and also allows users to select from several criteria on how duplicates in retrieved information will be eliminated. Efficient detection and elimination of duplicate database records and duplicate retrievals by search engines, which are very similar but not necessarily identical, have been investigated extensively by many scientists, e.g., Hernandez [1996]; Hernandez and Stolfo [1995]; Hylton [1996]; Monge and Elkan [1998]; and Silberschatz et al. [1995].

Card et al. [1996] developed two 3D virtual interface tools, WebBook and WebForager, for browsing and recording Web pages. Kobayashi et al. [1999] developed a system to compare how relevance ranking of documents differ when queries are changed. The *parallel ranking* system can be used in a variety of applications, e.g., query refinement and understanding the contents of a database from different perspectives (each query represents a different user perspective). Manber et al. [1997] developed WebGlimpse, a tool for simultaneous searching and browsing Web pages, which is based on the Glimpse search engine.

Morohashi et al. [1995] and Takeda and Nomiyama [1997] developed a system that uses new technologies to organize and display, in an easily discernible form, a massive set of data. The system, called "*information outlining*," extracts and analyzes a variety of features of the data set and interactively visualizes these features through corresponding multiple, graphical viewers. Interactions with multiple viewers facilitates reducing candidate results, profiling information, and discovering new facts. Sakairi [1999] developed a site map for visualizing a Web site's structure and keywords.

2.3.3 *Acoustical Interfaces*. Web-based IR contributes to the acceleration of studies on and development of more user friendly, nonvisual, input-output interfaces. Some examples of research projects are given in a special journal issue on the topic "*the next generation graphics user interfaces (GUIs)*" [CACM 1993]. An article in *Business Week* [1977] discusses user preference for speech-based interfaces, i.e., spoken input (which relies on speech recognition technologies) and spoken output (which relies on text-to-speech and speech synthesis technologies).

One response to this preference by Asakawa [1996] is a method to enable the visually impaired to access and use the Web interactively, even when Japanese and English appear on a page (IBM Homepage on Systems for the Disabled <trl.ibm.co.jp/projects/s7260/sysde.htm>). The basic idea is to identify different languages (e.g., English, Japanese) and different text types (e.g., title and section headers, regular text, hot buttons) and then assign persons with easily distinguishable voices (e.g., male, female) to read each of the different types of text. More recently, the method has been extended to enable the visually impaired to access tables in HTML [Oogane and Asakawa 1998].

Another solution, developed by Raman [1996], is a system that enables visually impaired users to surf the Web interactively. The system, called *Emacspeak*, is much more sophisticated than screen readers. It reveals the structure of a document (e.g., tables or calendars) in addition to reading the text aloud.

A third acoustic-based approach for Web browsing is being investigated by Mereu and Kazman [1996]. They examined how sound environments can be used for navigation and found that sighted users prefer musical environments to enhance conventional means of navigation, while the visually impaired prefer the use of tones. The components

of all of the systems described above can be modified for more general systems (i.e., not necessarily for the visually impaired) which require an audio/speech-based interface.

2.4 Ranking Algorithms for Web-Based Searches

A variety of techniques have been developed for ranking retrieved documents for a given input query. In this section we give references to some classical techniques that can be modified for use by Web search engines [Baeza-Yates and Ribeiro-Neto 1999; Berry and Browne 1999; Frakes and Baeza-Yates 1992]. Techniques developed specifically for the Web are also presented.

Detailed information regarding ranking algorithms used by major search engines is not publicly available, however—it seems that most use term weighting or variations thereof or vector space models [Baeza-Yates and Ribeiro-Neto 1999]. In vector space models, each document (in the database under consideration) is modeled by a vector, each coordinate of which represents an attribute of the document [Salton 1971]. Ideally, only those that can help to distinguish documents are incorporated in the attribute space. In a Boolean model, each coordinate of the vector is zero (when the corresponding attribute is absent) or unity (when the corresponding attribute is present). Many refinements of the Boolean model exist. The most commonly used are term-weighting models, which take into account the frequency of appearance of an attribute (e.g., keyword) or location of appearance (e.g., keyword in the title, section header, or abstract). In the simplest retrieval and ranking systems, each query is also modeled by a vector in the same manner as the documents. The ranking of a document with respect to a query is determined by its "*distance*" to the query vector. A frequently used yardstick is the angle defined by a query and document vector.[3] Ranking a document is based on computation of the angle defined by the query and document vector. It is impractical for very large databases.

One of the more widely used vector space model-based algorithms for reducing the dimension of the document ranking problem is *latent semantic indexing* (LSI) [Deerwester et al. 1990]. LSI reduces the retrieval and ranking problem to one of significantly lower dimensions, so that retrieval from very large databases can be performed in real time. Although a variety of algorithms based on document vector models for clustering to expedite retrieval and ranking have been proposed, LSI is one of the few that successfully takes into account *synonymy* and *polysemy*. Synonymy refers to the existence of equivalent or similar terms, which can be used to express an idea or object in most languages, and polysemy refers to the fact that some words have multiple, unrelated meanings. Absence of accounting for synonymy will lead to many small, disjoint clusters, some of which should actually be clustered together, while absence of accounting for polysemy can lead to clustering together of unrelated documents.

In LSI, documents are modeled by vectors in the same way as Salton's vector space model. We represent the relationship between the attributes and documents by an m-by-n (rectangular) matrix A, with ij-th entry a_{ij}, i.e.,

$$A = [a_{ij}].$$

The column vectors of A represent the documents in the database. Next, we compute the singular value decomposition (SVD) of A, then construct a modified matrix A_k, from the k largest singular

[3]The angle between two vectors is determined by computing the dot product and dividing by the product of the l_2-norms of the vectors.

values σ_i; $i = 1, 2, \ldots, k$, and their corresponding vectors, i.e.,

$$A_k = U_k \Sigma_k V_k^T.$$

Σ_k is a diagonal matrix with monotonically decreasing diagonal elements σ_i. U_k and V_k are matrices whose columns are the left and right singular vectors of the k largest singular values of A.[4]

Processing the query takes place in two steps: projection followed by matching. In the projection step, input queries are mapped to pseudodocuments in the reduced query-document space by the matrix U_k, then weighted by the corresponding singular values σ_i from the reduced rank singular matrix Σ_k. The process can be described mathematically as

$$q \rightarrow \hat{q} = q^T U_k \Sigma_k^{-1},$$

where q represents the original query vector; $\hat{q}$ the pseudodocument; q^T the transpose of q; and $(\cdot)^{-1}$ the inverse operator. In the second step, similarities between the pseudodocument $\hat{q}$ and documents in the reduced term document space V_k^T are computed using any one of many similarity measures, such as angles defined by each document and query vector; see Anderberg [1973] or Salton [1989]. Notable reviews of linear algebra techniques, including LSI and its applications to information retrieval, are Berry et al. [1995] and Letsche and Berry [1997].

Statistical approaches used in natural language modeling and IR can probably be extended for use by Web search engines. These approaches are reviewed in Crestani et al. [1998] and Manning and Schutze [1999].

Several scientists have proposed information retrieval algorithms based on analysis of hyperlink structures for use on the Web [Botafogo et al. 1992; Carriere and Kazman 1997; Chakrabarti et al. 1988; Chakrabarti et al. 1998; Frisse 1988; Kleinberg 1998; Pirolli et al. 1996; and Rivlin et al. 1994].

A simple means to measure the quality of a Web page, proposed by Carriere and Kazman [1997], is to count the number of pages with pointers to the page, and is used in the WebQuery system and the Rankdex search engine <rankdex.gari.com>. Google, which currently indexes about 85 million Web pages, is another search engine that uses link infomation. Its rankings are based, in part, on the number of other pages with pointers to the page. This policy seems to slightly favor educational and government sites over commercial ones. In November 1999, Northern Light introduced a new ranking system, which is also based, in part, on link data (Search Engine Briefs <searchenginewatch.com/sereport/99/11-briefs.html>).

The hyperlink structures are used to rank retrieved pages, and can also be used for clustering relevant pages on different topics. This concept of coreferencing as a means of discovering so-called "*communities*" of good works was originally introduced in nonInternet-based studies on cocitations by Small [1973] and White and McCain [1989].

Kleinberg [1998] developed an algorithm to find the several most information-rich or, *authority*, pages for a query. The algorithm also finds *hub* pages, i.e., pages with links to many *authority* pages, and labels the two types of retrieved pages appropriately.

3. FUTURE DIRECTIONS

In this section we present some promising and imaginative research endeavors that are likely to make an impact on Web use in some form or variation in the future. Knowledge management [IEEE 1998b].

[4]For details on implementation of the SVD algorithm, see Demmel [1997]; Golub and Loan [1996]; and Parlett [1998].

3.1 Intelligent and Adaptive Web Services

As mentioned earlier, research and development of *intelligent agents* (also known as *bots*, *robots*, and aglets) for performing specific tasks on the Web has become very active [Finin et al. 1998; IEEE 1996a]. These agents can tackle problems including finding and filtering information; customizing information; and automating completion of simple tasks [Gilbert 1997]. The agents "gather information or perform some other service without (the user's) immediate presence and on some regular schedule" (whatis?com home page <whatis.com/intellig.htm>). The BotSpot home page <botspot.com> summarizes and points to some historical information as well as current work on intelligent agents. The Proceedings of the Association for Computing Machinery (ACM), see Section 5.1 for the URL; the Conferences on Information and Knowledge Management (CIKM); and the American Association for Artificial Intelligence Workshops <www.aaai.org> are valuable information sources. The Proceedings of the Practical Applications of Intelligent Agents and Multi-Agents (PAAM) conference series <demon.co.uk/ar/paam96>and<demon.co.uk/ar/paam97> gives a nice overview of application areas. The home page of the IBM Intelligent Agent Center of Competence (IACC) <networking.ibm.com/iag/iaghome.html> describes some of the company's commercial agent products and technologies for the Web.

Adaptive Web services is one interesting area in intelligent Web robot research, including, e.g., Ahoy! The Homepage Finder, which performs dynamic reference sifting [Shakes et al. 1997]; Adaptive Web Sites, which "automatically improve their organization and presentation based on user access data" [Etzioni and Weld 1995; Perkowitz and Etzioni 1999]; Perkowitz's home page <info.cs.vt.edu>; and Adaptive Web Page Recommendation Service [Balabanovic 1997; Balabanovic and Shoham 1998; Balabanovic et al. 1995]. Discussion and ratings of some of these and other robots are available at several Web sites, e.g., Felt and Scales <wsulibs.wsu.edu/general/robots.htm> and Mitchell [1998].

Some scientists have studied prototype *metasearchers*, i.e., services that combine the power of several search engines to search a broader range of pages (since any given search engine covers less than 16% of the Web) [Gravano 1997; Lawrence and Giles 1998a; Selberg and Etzioni 1995a; 1995b]. Some of the better known metasearch engines include MetaCrawler, SavvySearch, and InfoSeek Express. After a query is issued, metasearchers work in three main steps: first, they evaluate which search engines are likely to yield valuable, fruitful responses to the query; next, they submit the query to search engines with high ratings; and finally, they merge the retrieved results from the different search engines used in the previous step. Since different search engines use different algorithms, which may not be publicly available, ranking of merged results may be a very difficult task.

Scientists have investigated a number of approaches to overcome this problem. In one system, a result merging condition is used by a metasearcher to decide how much data will be retrieved from each of the search engine results, so that the top objects can be extracted from search engines without examining the entire contents of each candidate object [Gravano 1997]. Inquirus downloads and analyzes individual documents to take into account factors such as query term context, identification of dead pages and links, and identification of duplicate (and near duplicate) pages [Lawrence and Giles 1998a]. Document ranking is based on the downloaded document itself, instead of rankings from individual search engines.

3.2 Information Retrieval for Internet Shopping

An intriguing application of Web robot technology is in simulation and prediction

of pricing strategies for sales over the Internet. The 1999 Christmas and holiday season marked the first time that shopping online was no longer a prediction; "Online sales increased by 300 percent and the number of orders increased by 270 percent" compared to the previous year [Clark 2000]. To underscore the point, *Time* magazine selected Jeff Bezos, founder of Amazon.com as 1999 Person of the Year. Exponential growth is predicted in online shopping. Charts that illustrate projected growth in Internet-generated revenue, Internet-related consumer spending, Web advertising revenue, etc. from the present to 2002, 2003, and 2005 are given in Nua's survey pages (see Section 1.2 for the URL).

Robots to help consumers shop, or *shopbots*, have become commonplace in e-commerce sites and general-purpose Web portals. Shopbot technology has taken enormous strides since its initial introduction in 1995 by Anderson Consulting. This first bot, known as Bargain Finder, helped consumers find the lowest priced CDs. Many current shopbots are capable of a host of other tasks in addition to comparing prices, such as comparing product features, user reviews, delivery options, and warranty information. Clark [2000] reviews the state-of-the-art in bot technology and presents some predicitions for the future by experts in the field—for example, Kephart, manager of IBM's Agents and Emergent Phenomena Group, predicts that "shopping bots may soon be able to negotiate and otherwise work with vendor bots, interacting via ontologies and distributed technologies... bots would then become 'economic actors making decisions'" and Guttman, chief technology officer for Frictionless commerce <frictionless.com> footnotes that Frictionless's bot engine is used by some famous portals, including Lycos, and mentions that his company's technology will be used in a retailer bot that will "negotiate trade-offs between product price, performance, and delivery times with shopbots on the basis of customer preferences." Price comparison robots and their possible roles in Internet merchant price wars in the future are discussed in Kephart et al. [1998a; 1998b].

The auction site is another successful technological off-shoot of the Internet shopping business [Cohen 2000; Ferguson 2000]. Two of the more famous general online auction sites are priceline.com <priceline.com> and eBay <ebay.com>. Priceline.com pioneered and patented its business concept, i.e., online bidding [Walker et al. 1997]. Patents related to that of priceline.com include those owned by ADT Automotive, Inc. [Berent et al. 1998]; Walker Asset Management [Walker et al. 1996]; and two individuals [Barzilai and Davidson 1997].

3.3 Multimedia Retrieval

IR from multimedia databases is a multidisciplinary research area, which includes topics from a very diverse range, such as analysis of text, image and video, speech, and nonspeech audio; graphics; animation; artificial intelligence; human-computer interaction; and multimedia computing [Faloutsos 1996; Faloutsos and Lin 1995; Maybury 1997; and Schauble 1997]. Recently, several commercial systems that integrate search capabilities from multiple databases containing heterogeneous, multimedia data have become available. Examples include PLS <pls.com>; Lexis-Nexis <lexis-nexis.com>; DIALOG <dialog.com>; and Verity <verity.com>. In this section we point to some recent developments in the field; but the discussion is by no means comprehensive.

Query and retrieval of images is one of the more established fields of research involving multimedia databases [IEEE ICIP: *Proceedings of the IEEE International Conference on Image Processing* and IEEE ICASSP: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* and IFIP 1992]. So much work by so many has been conducted on this topic that a comprehensive review is beyond the scope of this paper. But some se-

lected work in this area follows: search and retrieval from large image archives [Castelli et al. 1998]; pictorial queries by image similarity [Soffer and Samet]; image queries using Gabor wavelet features [Manjunath and Ma 1996]; fast, multiresolution image queries using Haar wavelet transform coefficients [Jacobs et al. 1995]; acquisition, storage, indexing, and retrieval of map images [Samet and Soffer 1986]; real-time fingerprint matching from a very large database [Ratha et al. 1992]; querying and retrieval using partially decoded JPEG data and keys [Schneier and Abdel-Mottaleb 1996]; and retrieval of faces from a database [Bach et al. 1993; Wu and Narasimhalu 1994].

Finding documents that have images of interest is a much more sophisticated problem. Two well-known portals with a search interface for a database of images are the Yahoo! Image Surfer <isurf.yahoo.com> and the Alta Vista PhotoFinder <image.altavista.com>. Like Yahoo!'s text-based search engine, the Image Surfer home pages are organized into categories. For a text-based query, a maximum of six thumbnails of the top-ranked retrieved images are displayed at a time, along with their titles. If more than six are retrieved, then links to subsequent pages with lower relevance rankings appear at the bottom of the page. The number of entries in the database seem to be small; we attempted to retrieve photos of some famous movie stars and came up with none (for Brad Pitt) or few retrievals (for Gwyneth Paltrow), some of which were outdated or unrelated links. The input interface to Photofinder looks very much like the interface for Alta Vista's text-based search engine. For a text-based query, a maximum of twelve thumbnails of retrieved images are displayed at a time. Only the name of the image file is displayed, e.g., *image.jpg*. To read the description of an image (if it is given), the mouse must point to the corresponding thumbnail. The number of retrievals for Photofinder were huge (4232 for Brad Pitt and 119 for Gwyneth Paltrow), but there was a considerable amount of noise after the first page of retrievals and there were many redundancies. Other search engines with an option for searching for images in their advanced search page are Lycos, HotBot, and AltaVista. All did somewhat better than Photofinder in retrieving many images of Brad Pitt and Gwyneth Paltrow; most of the thumbnails were relevant for the first several pages (each page contained 10 thumbnails).

NEC's Inquirus is an image search engine that uses results from several search engines. It analyzes the text accompanying images to determine relevance for ranking, and downloads the actual images to create thumbnails that are displayed to the user [Lawrence and Giles 1999c].

Query and retrieval of images in a video frame or frames is a research area closely related to retrieval of still images from a very large image database [Bolle et al. 1998]. We mention a few to illustrate the potentially wide scope of applications, e.g., content-based video indexing retrieval [Smoliar and Zhang 1994]; the Query-by-Image-Content (QBIC) system, which helps users find still images in large image and video databases on the basis of color, shape, texture, and sketches [Flickner et al. 1997; Niblack 1993]; Information Navigation System (INS) for multimedia data, a system for archiving and searching huge volumes of video data via Web browsers [Nomiyama et al. 1997]; and VisualSEEk, a tool for searching, browsing, and retrieving images, which allows users to query for images using the visual properties of regions and their spatial layout [Smith and Chang 1997a; 1996]; compressed domain image manipulation and feature extraction for compressed domain image and video indexing and searching [Chang 1995; Zhong and Chang 1997]; a method for extracting visual events from relatively long videos uing objects (rather than keywords), with specific applications to sports events [Iwai et al. 2000; Kurokawa et al. 1999]; retrieval and semantic

interpretation of video contents based on objects and their behavior [Echigo et al. 2000]; shape-based retrieval and its application to identity checks on fish [Schatz 1997]; and searching for images and videos on the Web [Smith and Chang 1997b].

Multilingual communication on the Web [Miyahara et al. 2000] and cross-language document retrieval is a timely research topic being investigated by many [Ballesteros and Croft 1998; Eichmann et al. 1998; Pirkola 1998]. An introduction to the subject is given in Oard [1997b], and some surveys are found in CLIR [1999] (Cross-Language Information Retrieval Project <clis.umd.edu/dlrg>); Oard [1997a] <glue.umd.edu/oard/research.html> and in Oard and Door [1996]. Several search engines now feature multilingual search, e.g., Open Text Web Index <index.opentext.net> searches in four languages (English, Japanese, Spanish, and Portuguese). A number of commercial Japanese-to-English and English-to-Japanese Web translation software products have been developed by leading Japanese companies in Japanese <bekkoame.ne.jp/oto3>. A typical example, which has a trial version for downloading, is a product called *Honyaku no Oosama* <ibm.co.jp/software/internet/king/index.html>, or Internet King of Translation [Watanabe and Takeda 1998].

Other interesting research topics and applications in multimedia IR are speech-based IR for digital libraries [Oard 1997c] and retrieval of songs from a database when a user hums the first few bars of a tune [Kageyama and Takashima 1994]. The melody retrieval technology has been incorporated as an interface in a karaoke machine.

3.4 Conclusions

Potentially lucrative application of Internet-based IR is a widely studied and hotly debated topic. Some pessimists believe that current rates of increase in the use of the Internet, number of Web sites and hosts are not sustainable, so that research and business opportunities in the area will decline. They cite statistics such as the April 1998 GVU WWW survey, which states that the use of better equipment (e.g., upgrades in modems by 48% of people using the Web) has not resolved the problem of slow access, and an August 1998 survey by Alexa Internet stating that 90% of all Web traffic is spread over 100,000 different hosts, with 50% of all Web traffic headed towards the top 900 most popular sites. In short, the pessimists maintain that an effective means of managing the highly uneven concentration of information packets on the Internet is not immediately available, nor will it be in the near future. Furthermore, they note that the exponential increase in Web sites and information on the Web is contributing to the second most commonly cited problem, that is, users not being able to find the information they seek in a simple and timely manner.

The vast majority of publications, however, support a very optimistic view. The visions and research projects of many talented scientists point towards finding concrete solutions and building more efficient and user-friendly solutions. For example, McKnight and Boroumand [2000] maintain that flat rate Internet retail pricing—currently the predominant pricing model in the U.S.—may be one of the major culprits in the traffic-congestion problem, and they suggest that other pricing models are being proposed by researchers. It is likely that the better proposals will be seriously considered by the business community and governments to avoid the continuation of the current solution, i.e., overprovisioning of bandwidth.

ACKNOWLEDGMENTS

The authors acknowledge helpful conversations with Stuart McDonald of alphaWorks and our colleagues at IBM Research. Our manuscript has benefitted greatly from the extensive and well-documented list of suggestions and corrections

from the reviewers of the first draft. We appreciate their generosity, patience, and thoughtfulness.

REFERENCES

ASSOCIATION FOR COMPUTING MACHINERY. 2000. SIGCHI: Special Interest Group on Computer-Human Interaction. Home page: www.acm.org/sigchi/

ASSOCIATION FOR COMPUTING MACHINERY. 2000. SIGIR: Special Interest Group on Information Retrieval. Home page: www.acm.org/sigir/

AGOSTI, M. AND SMEATON, A. 1996. *Information Retrieval and Hypertext*. Kluwer Academic Publishers, Hingham, MA.

AGRAWAL, R., GEHRKE, J., GUNOPULOS, D., AND RAGHAVAN, P. 1998. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (SIGMOD, Seattle, WA, June). ACM Press, New York, NY, 94–105.

AHLBERG, C. AND SHNEIDERMAN, B. 1994. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of the ACM Conference on Human Factors in Computing Systems: Celebrating Interdependence* (CHI '94, Boston, MA, Apr. 24–28). ACM Press, New York, NY, 313–317.

AHLBERG, C. AND SHNEIDERMAN, B. 1997. The alphaslider: A compact and rapid and selector. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '97, Atlanta, GA, Mar. 22–27), S. Pemberton, Ed. ACM Press, New York, NY.

AI MAG. 1997. Special issue on intelligent systems on the internet. *AI Mag. 18*, 4.

ANDERBERG, M. R. 1973. *Cluster Analysis for Applications*. Academic Press, Inc., New York, NY.

ANICK, P. G. AND VAITHYANATHAN, S. 1997. Exploiting clustering and phrases for context-based information retrieval. *SIGIR Forum 31*, 1, 314–323.

ASAKAWA, C. 1996. Enabling the visually disabled to use the www in a gui environment. IEICE Tech. Rep. HC96-29.

BACH, J., PAUL, S., AND JAIN, R. 1993. A visual information management system for the interactive retrieval of faces. *IEEE Trans. Knowl. Data Eng. 5*, 4, 619–628.

BAEZA-YATES, R. A. 1992. Introduction to data structures and algorithms related to information retrieval. In *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 13–27.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*. Addison-Wesley, Reading, MA.

BALABANOVIC, M. 1997. An adaptive Web page recommendation service. In *Proceedings of the First International Conference on Autonomous Agents* (AGENTS '97, Marina del Rey, CA, Feb. 5–8), W. L. Johnson, Chair. ACM Press, New York, NY, 378–385.

BALABANOVIC, M. AND SHOHAM, Y. 1995. Learning information retrieval agents: Experiments with automated web browsing. In *Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogenous Distributed Environments* (Stanford, CA, Mar.). AAAI Press, Menlo Park, CA.

BALABANOVIC, M., SHOHAM, Y., AND YUN, T. 1995. An adaptive agent for automated web browsing. Stanford Univ. Digital Libraries Project, working paper 1995-0023. Stanford University, Stanford, CA.

BALDONADO, M. 1997. An interactive, structure-mediated approach to exploring information in a heterogeneous, distributed environment. Ph.D. Dissertation. Computer Systems Laboratory, Stanford Univ., Stanford, CA.

BALDONADO, M. Q. W. AND WINOGRAD, T. 1997. SenseMarker: an information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '97, Atlanta, GA, Mar. 22–27), S. Pemberton, Ed. ACM Press, New York, NY, 11–18.

BALLESTEROS, L. AND CROFT, W. B. 1998. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '98, Melbourne, Australia, Aug. 24–28), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Chairs. ACM Press, New York, NY, 64–71.

BARZILAI AND DAVIDSON. 1997. Computer-based electronic bid, auction and sale system, and a system to teach new/non-registered customers how bidding, auction purchasing works: U.S. Patent no. 60112045.

BEAUDOIN, L., PARENT, M.-A., AND VROOMEN, L. C. 1996. Cheops: A compact explorer for complex hierarchies. In *Proceedings of the IEEE Conference on Visualization* (San Francisco, CA, Oct. 27-Nov. 1), R. Yagel and G. M. Nielson, Eds. IEEE Computer Society Press, Los Alamitos, CA, 87ff.

BEDERSON, B. B. AND HOLLAN, J. D. 1994. Pad++: A zooming graphical interface for exploring alternate interface physics. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology* (UIST '94, Marina del Rey, CA, Nov. 2–4), P. Szekely, Chair. ACM Press, New York, NY, 17–26.

BERENT, T., HURST, D., PATTON, T., TABERNIK, T., REIG, J. W. D., AND WHITTLE, W. 1998. Electronic on-line motor vehicle auction and information system: U.S. Patent no. 5774873.

BERNERS-LEE, T., CAILLIAU, R., LUOTONEN, A., NIELSEN, H. F., AND SECRET, A. 1994. The

World-Wide Web. *Commun. ACM 37*, 8 (Aug.), 76–82.

BERRY, M. AND BROWN, M. 1999. *Understanding Search Engines*. SIAM, Philadelphia, PA.

BERRY, M. W., DUMAIS, S. T., AND O'BRIEN, G. W. 1995. Using linear algebra for intelligent information retrieval. *SIAM Rev. 37*, 4 (Dec.), 573–595.

BHARAT, K. AND BRODER, A. 1998. A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of the Seventh International Conference on World Wide Web 7* (WWW7, Brisbane, Australia, Apr. 14–18), P. H. Enslow and A. Ellis, Eds. Elsevier Sci. Pub. B. V., Amsterdam, The Netherlands, 379–388.

BOLLE, R., YEO, B.-L., AND YEUNG, M. 1998. Video query: Research directions. *IBM J. Res. Dev. 42*, 2 (Mar.), 233–251.

BORKO, H. 1979. Inter-indexer consistency. In *Proceedings of the Cranfield Conference*.

BOTAFOGO, R. A., RIVLIN, E., AND SHNEIDERMAN, B. 1992. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Trans. Inf. Syst. 10*, 2 (Apr.), 142–180.

BRAKE, D. 1997. Lost in cyberspace. *New Sci. Mag.* www.newscientist.com/keysites/networld/lost.html

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst. 30*, 1-7, 107–117.

BRODER, A. Z., GLASSMAN, S. C., MANASSE, M. S., AND ZWEIG, G. 1997. Syntactic clustering of the Web. *Comput. Netw. ISDN Syst. 29*, 8-13, 1157–1166.

BUSINESS WEEK. 1997. Special report on speech technologies. *Business Week*.

COMMUNICATIONS OF THE ACM. 1993. Special issue on the next generation GUIs. *Commun. ACM*.

COMMUNICATIONS OF THE ACM. 1994. Special issue on internet technology. *Commun. ACM*.

COMMUNICATIONS OF THE ACM. 1995. Special issues on digital libraries. *Commun. ACM*.

COMMUNICATIONS OF THE ACM. 1999. Special issues on knowledge discovery. *Commun. ACM*.

CARD, S., MACKINLAY, J., AND SHNEIDERMAN, B. 1999. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

CARD, S. K., ROBERTSON, G. G., AND YORK, W. 1996. The WebBook and the Web Forager: an information workspace for the World-Wide Web. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '96, Vancouver, B.C., Apr. 13–18), M. J. Tauber, Ed. ACM Press, New York, NY, 111ff.

CARL, J. 1995. Protocol gives sites way to keep out the 'bots'. *Web Week 1*, 7 (Nov.).

CARRIÈRE, J. AND KAZMAN, R. 1997. WebQuery: Searching and visualizing the Web through connectivity. In *Proceedings of the Sixth International Conference on the World Wide Web* (Santa Clara CA, Apr.).

CASTELLI, V., BERGMAN, L., KONTOYIANNINS, I., LI, C.-S., ROBINSON, J., AND TUREK, J. 1998. Progressive search and retrieval in large image archives. *IBM J. Res. Dev. 42*, 2 (Mar.), 253–268.

CATHRO, W. 1997. Matching discovery and recovery. In *Proceedings of the Seminar on Standards Australia*. www.nla.gov.au/staffpaper/cathro3.html

CHAKRABARTI, S., DOM, B., GIBSON, D., KUMAR, S., RAGHAVAN, P., RAJAGOPALAN,, S., AND TOMKINS, A. 1988. Experiments in topic distillation. In *Proceedings of the ACM SIGIR Workshop on Hypertext Information Retrieval for the Web* (Apr.). ACM Press, New York, NY.

CHAKRABARTI, S., DOM, B., RAGHAVAN, P., RAJAGOPALAN, S., GIBSON, D., AND KLEINBERG, J. 1998. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the Seventh International Conference on World Wide Web 7* (WWW7, Brisbane, Australia, Apr. 14–18), P. H. Enslow and A. Ellis, Eds. Elsevier Sci. Pub. B. V., Amsterdam, The Netherlands, 65–74.

CHAKRABARTI, S. AND RAJAGOPALAN, S. 1997. Survey of information retrieval research and products. Home page: w3.almaden.ibm.com/soumen/ir.html

CHALMERS, M. AND CHITSON, P. 1992. Bead: explorations in information visualization. In *Proceedings of the 15th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR '92, Copenhagen, Denmark, June 21–24), N. Belkin, P. Ingwersen, and A. M. Pejtersen, Eds. ACM Press, New York, NY, 330–337.

CHANDRASEKARAN, R. 1998. "Portals" offer one-stop surfing on the net. *Int. Herald Tribune 19/21*.

CHANG, S.-F. 1995. Compressed domain techniques for image/video indexing and manipulation. In *Proceedings of the Conference on Information Processing*.

CHO, J., GARCIA-MOLINA, H., AND PAGE, L. 1998. Efficient crawling through URL ordering. *Comput. Netw. ISDN Syst. 30*, 1-7, 161–172.

CLARK, D. 2000. Shopbots become agents for business change. *IEEE Computer*.

CLEVERDON, C. 1970. Progress in documentation. *J. Doc. 26*, 55–67.

CLIR. 1999. Cross-language information retrieval project, resource page. Tech. Rep. University of Maryland at College Park, College Park, MD.

COHEN, A. 1999. The attic of e. *Time Mag.*

COMPUT. NETW. ISDN SYST. 2000. World Wide Web conferences. 1995-2000. *Comput. Netw. ISDN Syst.* www.w3.org/Conferences/Overview-WWW.html

COOPER, W. 1969. Is interindexer consistency a hobgoblin? *Am. Doc. 20*, 3, 268–278.

CRANOR, L. F. AND LA MACCHIA, B. A. 1998. Spam! *Commun. ACM 41*, 8, 74–83.

CRESTANI, F., LALMAS, M., VAN RIJSBERGEN, C. J., AND CAMPBELL, I. 1998. Is this document relevant?

Probably: A survey of probabilistic models in information retrieval. *ACM Comput. Surv. 30*, 4, 528–552.

CUNNINGHAM, M. 1997. Brewster's millions. *Irish Times*.www.irish-times.com/irish-times/paper/1997/0127/cmp1.html

CUTTING, D. R., KARGER, D. R., AND PEDERSEN, J. O. 1993. Constant interaction-time scatter/gather browsing of very large document collections. In *Proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR '93, Pittsburgh, PA, June 27–July), R. Korfhage, E. Rasmussen, and P. Willett, Eds. ACM Press, New York, NY, 126–134.

DEERWESTER, S., DUMAI, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci. 41*, 6, 391–407.

DEMMEL, J. W. 1997. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA.

DHILLON, I. AND MOHDA, D. 1999. A data-clustering algorithm on distributed memory multiprocessors. In *Proceedings of the Workshop on Large-Scale Parallel KDD Systems* (ACM SIGKDD., Aug. 15-18). ACM Press, New York, NY.

DHILLON, I. AND MODHA, D. 2000. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*

ECHIGO, T., KUROKAWA, M., TOMITA, A., TOMITA, A., MIYAMORI, AND IISAKU, S. 2000. Video enrichment: Retrieval and enhanced visualization based on behaviors of objects. In *Proceedings of the Fourth Asian Conference on Computer Vision* (ACCV2000, Jan. 8-11). 364–369.

EICHMANN, D., RUIZ, M. E., AND SRINIVASAN, P. 1998. Cross-language information retrieval with the UMLS metathesaurus. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '98, Melbourne, Australia, Aug. 24–28), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Chairs. ACM Press, New York, NY, 72–80.

ESTER, M., KRIEGEL, H.-S., SANDER, J., AND XU, X. 1995a. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (Montreal, Canada, Aug. 20-21).

ESTER, M., KRIEGEL, H.-S., AND XU, X. 1995b. A database interface for clustering in large spatial databases. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (Montreal, Canada, Aug. 20-21).

ESTER, M., KRIEGEL, H.-S., AND XU, X. 1995c. Focusing techniques for efficient class identification. In *Proceedings of the Fourth International Symposium on Large Spatial Databases*.

ETZIONI, O. AND WELD, D. 1995. Intelligent agents on the Internet: Fact, fiction and forecast. Tech. Rep. University of Washington, Seattle, WA.

FALOUTSOS, C. 1996. *Searching Multimedia Databases by Content*. Kluwer Academic Publishers, Hingham, MA.

FALOUTSOS, C. AND LIN, K. 1995. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (ACM-SIGMOD, San Jose, CA, May). SIGMOD. ACM Press, New York, NY, 163–174.

FALOUTSOS, C. AND OARD, D. W. 1995. A survey of information retrieval and filtering methods. Univ. of Maryland Institute for Advanced Computer Studies Report. University of Maryland at College Park, College Park, MD.

FELDMAN, S. 1998. Web search services in 1998: Trends and challenges. *Inf. Today 9*.

FERGUSON, A. 1999. Auction nation. *Time Mag.*

FININ, T., NICHOLAS, C., AND MAYFIELD, J. 1998. Software agents for information retrieval (short course notes). In *Proceedings of the Third ACM Conference on Digital Libraries* (DL '98, Pittsburgh, PA, June 23–26), I. Witten, R. Akscyn, and F. M. Shipman, Eds. ACM Press, New York, NY.

FINKELSTEIN, A. AND SALESIN, D. 1995. Fast multiresolution image querying. In *Proceedings of the ACM SIGGRAPH Conference on Visualization: Art and Interdisciplinary Programs* (SIGGRAPH '95, Los Angeles, CA, Aug. 6–11), K. O'Connell, Ed. ACM Press, New York, NY, 277–286.

FISHER, D. 1995. Iterative optimization and simplification of hierarchical clusterings. Tech. Rep. Vanderbilt University, Nashville, TN.

FLICKNER, M., SAWHNEY, H., NIBLACK, W., ASHLEY, J., HUANG, Q., DOM, B., GORKANI, M., HAFNER, J., LEE, D., PETKOVIC, D., STEELE, D., AND YANKER, P. 1997. Query by image and video content: the QBIC system. In *Intelligent Multimedia Information Retrieval*, M. T. Maybury, Ed. MIT Press, Cambridge, MA, 7–22.

FLYNN, L. 1996. Desperately seeking surfers: Web programmers try to alter search engines' results. *New York Times*.

FRAKES, W. B. AND BAEZA-YATES, R., EDS. 1992. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Inc., Upper Saddle River, NJ.

FRISSE, M. E. 1988. Searching for information in a hypertext medical handbook. *Commun. ACM 31*, 7 (July), 880–886.

GILBERT, D. 1997. Intelligent agents: The right information at the right time. Tech. Rep. IBM Corp., Research Triangle Park, NC.

GLOOR, P. AND DYNES, S. 1998. Cybermap: Visually navigating the web. *J. Visual Lang. Comput. 9*, 3 (June), 319–336.

GOLUB, G. H. AND VAN LOAN, C. F. 1996. *Matrix Computations*. 3rd. Johns Hopkins studies in

the mathematical sciences. Johns Hopkins University Press, Baltimore, MD.

GRAVANO, L. 1998. Querying multiple document collections across the Internet. Ph.D. Dissertation. Stanford University, Stanford, CA.

GUDIVADA, V., RAGHAVAN, V., GROSKY, W., AND KASAANAGOTTU, R. 1997. Information retrieval on the world wide web. *IEEE Internet Comput. 1*, 1 (May/June), 58–68.

GUGLIELMO, C. 1997. Upside today (on-line). home page: inc.com/cgi-bin/tech link.cgi?url=http://www.upside.com.

GUHA, S., RASTOGI, R., AND SHIM, K. 1998. Cure: An efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (SIGMOD, Seattle, WA, June). ACM Press, New York, NY.

HAWKING, D., CRASWELL, N., THISTLEWAITE, P., AND HARMAN, D. 1999. *Results and Challenges in Web Search Evaluation*.

HEARST, M. A. 1995. TileBars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '95, Denver, CO, May 7–11), I. R. Katz, R. Mack, L. Marks, M. B. Rosson, and J. Nielsen, Eds. ACM Press/Addison-Wesley Publ. Co., New York, NY, 59–66.

HEARST, M. 1997. Interfaces for searching the web. *Sci. Am.*, 68–72.

HEARST, M. 1999. User interfaces and visualization. In *Modern Information Retrieval*, R. Baeza-Yates and B. Ribeiro-Neto. Addison-Wesley, Reading, MA, 2257–3232.

HEARST, M. A. AND PEDERSEN, J. O. 1996. Visualizing information retrieval results: a demonstration of the TileBar interface. In *Proceedings of the CHI '96 Conference Companion on Human Factors in Computing Systems: Common Ground* (CHI '96, Vancouver, British Columbia, Canada, Apr. 13–18), M. J. Tauber, Ed. ACM Press, New York, NY, 394–395.

HENZINGER, M., HEYDON, A., MITZENMACHER, M., AND NAJORK, M. 1999. Measuring index quality using random walks on the web.

HERNANDEZ, M. 1996. A generalization of band joins and the merge/purge problem. Ph.D. Dissertation. Columbia Univ., New York, NY.

HOWE, A. AND DREILINGER, D. 1997. Savvysearch: A metasearch engine that learns which search engine to query. *AI Mag. 18*, 2, 19–25.

HUBERMAN, B. AND LUKOSE, R. 1997. A metasearch engine that learns which search engine to query. *Science 277*, 535–537.

HUBERMAN, B., PIROLLI, P., PITKOW, J., AND LUKOSE, R. 1998. Strong regularities in world wide web surfing. *Science 280*, 95–97.

HYLTON, J. 1996. Identifying and merging related bibliographic records. Master's Thesis.

IEEE. 1999. Special issue on intelligent information retrieval. *IEEE Expert*.

IEEE. 1998a. News and trends section. *IEEE Internet Comput.*

IEEE. 1998b. Special issue on knowledge management. *IEEE Expert*.

IEEE. 1996a. Special issue on intelligent agents. *IEEE Expert/Intelligent Systems and Their Applications*.

IEEE. 1996b. Special issue on digital libraries: representation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 8, 771–859.

IFIP. 1989. *Visual Data Base Systems I and II*. Elsevier North-Holland, Inc., Amsterdam, The Netherlands.

IWAI, Y., MARUO, J., YACHIDA, M., ECHIGO, T., AND IISAKU, S. 2000. A framework for visual event extraction from soccer games. In *Proceedings of the Fourth Asian Conference on Computer Vision* (ACCV2000, Jan. 8-11). 222–227.

JACOBY, J. AND SLAMECKA, V. 1962. Indexer consistency under minimal conditions. RADC TR 62--426. Documentation, Inc., Bethesda, MD, US.

KAGEYAMA, T. AND TAKASHIMA, Y. 1994. A melody retrieval method with hummed melody. *IEICE Trans. Inf. Syst. J77*, 8 (Aug.), 1543–1551.

KAHLE, B. 1999. Archiving the Internet. home page: www.alexa.com/ brewster/essays/sciam article.html

KEPHART, J., HANSON, J., LEVINE, D., GROSOF, B., SAIRAMESH, J., AND WHITE, R. S. 1998a. Emergent behavior in information economies.

KEPHART, J. O., HANSON, J. E., AND SAIRAMESH, J. 1998b. Price-war dynamics in a free-market economy of software agents. In *Proceedings of the Sixth International Conference on Artificial Life* (ALIFE, Madison, WI, July 26–30), C. Adami, R. K. Belew, H. Kitano, and C. E. Taylor, Eds. MIT Press, Cambridge, MA, 53–62.

KLEINBERG, J. M. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the 1998 ACM-SIAM Symposium on Discrete Algorithms* (San Francisco CA, Jan.). ACM Press, New York, NY.

KOBAYASHI, M., DUPRET, G., KING, O., SAMUKAWA, H., AND TAKEDA, K. 1999. Multi-perspective retrieval, ranking and visualization of web data. In *Proceedings of the International Symposium on Digital Libraries* ((ISDL99), Tsukuba, Japan). 159–162.

KORFHAGE, R. R. 1997. *Information Storage and Retrieval*. John Wiley and Sons, Inc., New York, NY.

KOSTER, M. 1995. Robots in the web: trick or treat? *ConneXions 9*, 4 (Apr.).

KOSTER, M. 1996. Examination of the standard for robots exclusion. home page: info.webcrawler-.com/mak/projects/robots/eval.html

KUROKAWA, M., ECHIGO, T., TOMITA, T., MAEDA, J., MIYAMORI, H., AND ISISAKU, S. 1999. Representation and retrieval of video scene by using object actions and their spatio-temporal relationships. In *Proceedings of the Interna-*

tional Conference on ICIP-Image Processing. IEEE Press, Piscataway, NJ.

LAGOZE, C. 1996. The Warwick framework: A container architecture for diverse sets of metadata. *D-Lib Mag.* www.dlib.org

LAMPING, J., RAO, R., AND PIROLLI, P. 1995. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '95, Denver, CO, May 7–11), I. R. Katz, R. Mack, L. Marks, M. B. Rosson, and J. Nielsen, Eds. ACM Press/Addison-Wesley Publ. Co., New York, NY, 401–408.

LAWRENCE, S. AND GILES, C. 1998a. Context and page analysis for improved web search. *IEEE Internet Comput. 2*, 4, 38–46.

LAWRENCE, S. AND GILES, C. 1998b. Searching the world wide web. *Science 280*, 98–100.

LAWRENCE, S. AND GILES, C. 1999a. Accessibility of information on the web. *Nature 400*, 107–109.

LAWRENCE, S. AND GILES, C. 1999b. Searching the web: General and scientific information access. *IEEE Commun. Mag. 37*, 1, 116–122.

LAWRENCE, S. AND GILES, C. 1999c. Text and image metasearch on the web. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications* (PDPTA99). 829–835.

LEIGHTON, H. AND SRIVASTAVA, J. 1997. Precision among world wide web search engines: Altavista, excite, hotbot, infoseek, and lycos. home page: www.winona.msus.edu/library/webind2/webind2.htm.

LETSCHE, T. AND BERRY, M. 1997. Large-scale information retrieval with latent semantic indexing: (submitted). *Inf. Sci. Appl.*

LIAO, H., OSADA, M., AND SHNEIDERMAN, B. 1992. A formative evaluation of three interfaces for browsing directories using dynamic queries. Tech. Rep. CS-TR-2841. Department of Computer Science, University of Maryland, College Park, MD.

LIBERATORE, K. 1997. Getting to the source: Is it real or spam, ma'am ? *MacWorld*.

LIDSKY, D. AND KWON, R. 1997. Searching the net. *PC Mag.*, 227–258.

LIECHTI, O., SIFER, M. J., AND ICHIKAWA, T. 1998. Structured graph format: XML metadata for describing Web site structure. *Comput. Netw. ISDN Syst. 30*, 1-7, 11–21.

LOSEE, R. M. 1998. *Text Retrieval and Filtering: Analytic Models of Performance*. Kluwer international series on information retrieval. Kluwer Academic Publishers, Hingham, MA.

LYNCH, C. 1997. Searching the Internet. *Sci. Am.*, 52–56.

MAAREK, Y. S., JACOVI, M., SHTALHAIM, M., UR, S., ZERNIK, D., AND BEN-SHAUL, I. Z. 1997. WebCutter: A system for dynamic and tailorable site mapping. *Comput. Netw. ISDN Syst. 29*, 8-13, 1269–1279.

MACSKASSY, S., BANERJEE, A., DAVISON, B., AND HIRSH, H. 1998. Human performance on clustering web pages: A preliminary study. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (Seattle, WA, June '98). 264–268.

MANBER, U. 1999. Foreword. In *Modern Information Retrieval*, R. Baeza-Yates and B. Ribeiro-Neto. Addison-Wesley, Reading, MA, 5–8.

MANBER, U., SMITH, M., AND GOPAL, B. 1997. Webglimpse: Combining borwsing and searching. In *Proceedings on USENIX 1997 Annual Technical Conference* (Jan.). 195–206.

MANJUNATH, B. S. AND MA, W. Y. 1996. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 8, 837–842.

MANNING, C. AND SCHUTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

MARCHIONINI, G. 1995. *Information Seeking in Electronic Environments*. Cambridge Series on Human-Computer Interaction. Cambridge University Press, New York, NY.

MAYBURY, M. 1997. *Intelligent Multmedia Information Retrieval*. MIT Press, Cambridge, MA.

MAYBURY, M. T. AND WAHLSTER, W., EDS. 1998. *Readings in Intelligent User Interfaces*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

MCKNIGHT, L. 2000. Pricing internet services: Approaches and challenges. *IEEE Computer*, 128–129.

MEREU, S. W. AND KAZMAN, R. 1996. Audio enhanced 3D interfaces for visually impaired users. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '96, Vancouver, B.C., Apr. 13–18), M. J. Tauber, Ed. ACM Press, New York, NY, 72–78.

MITCHELL, S. 1998. General internet resource finding tools. home pages: library.ucr.edu/pubs/navigato.html

MIYAHARA, T, WATANABE, H., TAZOE, E., KAMIYAMA, Y., AND TAKEDA, K. 2000. *Internet Machine Translation*. Mainichi Communications, Japan.

MODHA, S. AND SPANGLER, W. 2000. Clustering hypertext with applications to web searching. In *Proceedings of the Conference on Hypertext* (May 30-June 3).

MONGE, A. AND ELKAN, C. 1998. An efficient domain-independent algorithm for detecting approximately duplicate database records. Tech. Rep. University of California at San Diego, La Jolla, CA.

MONIER, L. 1998. Altavista cto responds. www4.zdnet.com/anchordesk/talkback/talkback 13066.html.

MOROHASHI, M., TAKEDA, K., NOMIYAMA, H., AND MARUYAMA, H. 1995. Information outlining. In *Proceedings of International Symposium on Digital Libraries* (Tsukuba, Japan).

MUNZNER, T. AND BURCHARD, P. 1995. Visualizing the structure of the World Wide Web in 3D hyperbolic space. In *Procedings of the Symposium on Virtual Reality Modeling Language*

(VRML '95, San Diego, CA, Dec. 14–15), D. R. Nadeau and J. L. Moreland, Chairs. ACM Press, New York, NY, 33–38.

NAGAO, K. AND HASIDA, K. 1998. Automatic text summarization based on the global document annotation. In *Proceedings of the Conference on COLING-ACL*.

NAGAO, K., HOSOYA, S., KAWAKITA, Y., ARIGA, S., SHIRAI, Y., AND YURA, J. 1999. Semantic transcoding: Making the world wide web more understandable and reusable by external annotations.

NAVARRO, G. 1998. Approximate text searching. Ph.D. Dissertation. Univ. of Chile, Santiago, Chile.

NG, R. AND HAN, J. 1994. Efficient and effective methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases* (VLDB'94, Santiago, Chile, Sept.). VLDB Endowment, Berkeley, CA.

NIBLACK, W. 1993. The qbic project: Query by image by content using color, texture, and shape. In *Proceedings of the Conference on Storage and Retrieval for Image and Video Databases*. SPIE Press, Bellingham, WA, 173–187.

NIELSEN, J. 1993. *Usability Engineering*. Academic Press Prof., Inc., San Diego, CA.

NIELSEN, J. 1999. User interface directions for the Web. *Commun. ACM 42*, 1, 65–72.

NOMIYAMA, H., KUSHIDA, T., URAMOTO, N., IOKA, M., KUSABA, M., KUSABA, J.-K., CHIGONO, A., ITOH, T., AND TSUJI, M. 1997. Information navigation system for multimedia data. Res. Rep. RT-0227. Research Laboratory, IBM Tokyo, Tokyo, Japan.

OARD, D. 1997a. Cross-language text retrieval research in the USA. In *Proceedings of the Third Delos Workshop on ERCIM* (Mar.).

OARD, D. 1997b. Serving users in many languages. *D-Lib Mag. 3*, 1 (Jan.).

OARD, D. 1997c. Speech-based information retrieval for digital libraries. Tech. Rep. CS-TR-3778. University of Maryland at College Park, College Park, MD.

OARD, D. W. AND DORR, B. J. 1996. A survey of multilingual text retrieval. Tech. Rep UMIACS-TR-96-19. University of Maryland at College Park, College Park, MD.

OMIECINSKI, E. AND SCHEUERMANN, P. 1990. A parallel algorithm for record clustering. *ACM Trans. Database Syst. 15*, 4 (Dec.), 599–624.

OOGANE, T. AND ASAKAWA, C. 1998. An interactive method for accessing tables in HTML. In *Proceedings of the Third International ACM Conference on Assistive Technologies* (Assets '98, Marina del Rey, CA, Apr. 15–17), M. M. Blattner and A. I. Karshmer, Chairs. ACM Press, New York, NY, 126–128.

PARLETT, B. N. 1998. *The Symmetric Eigenvalue Problem*. Prientice-Hall SIAM Classics in Applied Mathematics Series. Prentice-Hall, Inc., Upper Saddle River, NJ.

PERKOWITZ, M. AND ETZIONI, O. 1999. Adaptive web sites: An ai challenge. Tech. Rep. University of Washington, Seattle, WA.

PIRKOLA, A. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '98, Melbourne, Australia, Aug. 24–28), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Chairs. ACM Press, New York, NY, 55–63.

PIROLLI, P., PITKOW, J., AND RAO, R. 1996. Silk from a sow's ear: extracting usable structures from the Web. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '96, Vancouver, B.C., Apr. 13–18), M. J. Tauber, Ed. ACM Press, New York, NY, 118–125.

PLAISANT, C. 1994. Dynamic queries on a health statistics atlas. Tech. Rep. University of Maryland at College Park, College Park, MD.

PRESCHEL, B. 1972. Indexer consistency in perception of concepts and choice of terminology. Final Rep. Columbia Univ., New York, NY.

RAGHAVAN, P. 1997. Information retrieval algorithms: A survey. In *Proceedings of the Symposium on Discrete Algorithms*. ACM Press, New York, NY.

RAMAN, T. V. 1996. Emacspeak—a speech interface. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '96, Vancouver, B.C., Apr. 13–18), M. J. Tauber, Ed. ACM Press, New York, NY, 66–71.

RAO, R., PEDERSEN, J. O., HEARST, M. A., MACKINLAY, J. D., CARD, S. K., MASINTER, L., HALVORSEN, P.-K., AND ROBERTSON, G. C. 1995. Rich interaction in the digital library. *Commun. ACM 38*, 4 (Apr.), 29–39.

RASMUSSEN, E. 1992. Clustering algorithms. In *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 419–442.

RATHA, N. K., KARU, K., CHEN, S., AND JAIN, A. K. 1996. A real-time matching system for large fingerprint databases. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 8, 799–813.

RENNISON, E. 1994. Galaxy of news: an approach to visualizing and understanding expansive news landscapes. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology* (UIST '94, Marina del Rey, CA, Nov. 2–4), P. Szekely, Chair. ACM Press, New York, NY, 3–12.

RIVLIN, E., BOTAFOGO, R., AND SHNEIDERMAN, B. 1994. Navigating in hyperspace: designing a structure-based toolbox. *Commun. ACM 37*, 2 (Feb.), 87–96.

ROBERTSON, G. G., MACKINLAY, J. D., AND CARD, S. K. 1991. Cone trees: Animated 3D visualizations of hierarchical information. In *Proceedings*

of the Conference on Human Factors in Computing Systems: Reaching through Technology (CHI '91, New Orleans, LA, Apr. 27–May 2), S. P. Robertson, G. M. Olson, and J. S. Olson, Eds. ACM Press, New York, NY, 189–194.

SAKAIRI, T. 1999. A site map for visualizing both a web site's structure and keywords. In *Proceedings of the IEEE Conference on System, Man, and Cybernetics* (SMC '99). IEEE Computer Society Press, Los Alamitos, CA, 200–205.

SALTON, G. 1969. A comparison between manual and automatic indexing methods. *Am. Doc. 20*, 1, 61–71.

SALTON, G. 1970. Automatic text analysis. *Science 168*, 335–343.

SALTON, G., ED. 1971. *The Smart Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, NJ.

SALTON, G., ED. 1988. *Automatic Text Processing*. Addison-Wesley Series in Computer Science. Addison-Wesley Longman Publ. Co., Inc., Reading, MA.

SALTON, G. AND MCGILL, M. J. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., Hightstown, NJ.

SAMET, H. AND SOFFER, A. 1996. MARCO: MAp Retrieval by COntent. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 8, 783–798.

SCHATZ, B. 1997. Information retrieval in digital libraries: Bringing search to the net. *Science 275*, 327–334.

SCHAUBLE, P. 1997. *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers, Hingham, MA.

SCIENTIFIC AMERICAN. 1997. The Internet: Fulfillling the promise: special report. Scientific American, Inc., New York, NY.

SELBERG, E. AND ETZIONI, O. 1995a. The metacrawler architecture for resource aggregation on the web. *IEEE Expert*.

SELBERG, E. AND ETZIONI, O. 1995b. Multiple service search and comparison using the metacrawler. In *Proceedings of the Fourth International Conference on The World Wide Web* (Boston, MA).

SHAKES, J., LANGHEINRICH, M., AND ETZIONI, O. 1997. Dynamic reference sifting: A case study in the homepage domain. In *Proceedings of the Conference on The World Wide Web*. 189–200.

SHIVAKUMAR, N. AND GARCIA-MOLINA, H. 1998. Finding near-replicas of documents on the web. In *Proceedings of the Workshop on Web Databases* (Valencia, Spain, Mar.).

SHNEIDERMAN, B. 1994. Dynamic queries for visual information seeking. Tech. Rep. CS-TR-3022. University of Maryland at College Park, College Park, MD.

SHNEIER, M. AND ABDEL-MOTTALEB, M. 1996. Exploiting the JPEG compression scheme for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell. 18*, 8, 849–853.

SILBERSCHATZ, A., STONEBRAKER, M., AND ULLMAN, J. 1995. Database research: Achievements and opportunities into the 21st century. In *Proceedings of the NSF Workshop on The Future of Database Research* (May).

SMALL, H. 1973. Co-citation in the scientific literature: a new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci. 24*, 265–269.

SMITH, J. R. AND CHANG, S.-F. 1996. VisualSEEk: A fully automated content-based image query system. In *Proceedings of the Fourth ACM International Conference on Multimedia* (Multimedia '96, Boston, MA, Nov. 18–22), P. Aigrain, W. Hall, T. D. C. Little, and V. M. Bove, Chairs. ACM Press, New York, NY, 87–98.

SMITH, J. R. AND CHANG, S.-F. 1997a. Querying by color regions using VisualSEEk content-based visual query system. In *Intelligent Multimedia Information Retrieval*, M. T. Maybury, Ed. MIT Press, Cambridge, MA, 23–41.

SMITH, J. AND CHANG, S.-F. 1997b. Searching for images and videos on the world-wide web. *IEEE MultiMedia*.

SMITH, Z. 1973. The truth about the web: Crawling towards eternity. *Web Tech. Mag.* www.webtechniques.com/features/1997/05/burner/burner.html

SMOLIAR, S. W. AND ZHANG, H. 1994. Content-based video indexing and retrieval. *IEEE MultiMedia 1*, 2 (Summer), 62–72.

SNEATH, P. H. A. AND SOKAL, R. R. 1973. *Numerical Taxonomy*. Freeman, London, UK.

SOERGEL, D. 1985. *Organizing Information: Principles of Data Base and Retrieval Systems*. Academic Press library and information science series. Academic Press Prof., Inc., San Diego, CA.

SOFFER, A. AND SAMET, H. 2000. Pictorial query specification for browsing through spatially-referenced image databases. *J. Visual Lang. Comput.*

SPARCK JONES, K. AND WILLETT, P., EDS. 1997. *Readings In Information Retrieval*. Morgan Kaufmann multimedia information and systems series. Morgan Kaufmann Publishers Inc., San Francisco, CA.

STOLFO, S. AND HERNANDEZ, M. 1995. The merge/purge problem for large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (San Jose, CA, May). 127–138.

STRATEGYALLEY. 1998. White paper on the viability of the internet for business. home page: www.strategyalley.com/articles/inet1.htm.

STRZALKOWSKI, T. 1999. *Natural Language Information Retreival*. Kluwer Academic Publishers, Hingham, MA.

TAKEDA, K. AND NOMIYAMA, H. 1997. Information outlining and site outlining. In *Proceedings of*

the International Symposium on Digital Libraries (ISDL97, Tsukuba, Japan).

TETRANET SOFTWARE INC. 1998. Wisebot. Home page for Wisebots: www.tetranetsoftware.com/products/wisebot.htm

TUFTE, E. R. 1986. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

VAN RIJSBERGEN, C. 1977. A theoretical basis for the use of cooccurrence data in information retrieval. *J. Doc. 33*, 2.

VAN RIJSBERGEN, C. 1979. *Information Retrieval*. 2nd ed. Butterworths, London, UK.

WALKER, J., CASE, T., JORASCH, J., AND SPARICO, T. 1996. Method, apparatus, and program for pricing, selling, and exercising options to purchase airline tickets: U.S. Patent no. 5797127.

WALKER, J., SPARICO, T., AND CASE, T. 1997. Method and apparatus for the sale of airline-specified flight tickets: U.S. Patent no. 5897620.

WATANABE, H. AND TAKEDA, K. 1998. A pattern-based machine translation system extended by example-based processing. In *Proceedings of the Conference on COLING-ACL*. 1369–1373.

WEBSTER, K. AND PAUL, K. 1996. Beyond surfing: Tools and techniques for searching the web. home page: magi.com/mmelick/it96jan.htm.

WESTERA, G. 1996. Robot-driven search engine evaluation overview. www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/.

WHITE, H. AND MCCAIN, K. 1989. *Bibliometrics. Annual Review Information Science and Technology*.

WILLETT, P. 1988. Recent trends in hierarchic document clustering: a critical review. *Inf. Process. Manage. 24*, 5 (), 577–597.

WILLIAMS, M. 1984. What makes rabbit run? *Int. J. Man-Mach. Stud. 2a*, 1, 333–352.

WILLIAMSON, C. AND SHNEIDERMAN, B. 1992. The dynamic HomeFinder: Evaluating dynamic queries in a real-estate information exploration system. In *Proceedings of the 15th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR '92, Copenhagen, Denmark, June 21–24), N. Belkin, P. Ingwersen, and A. M. Pejtersen, Eds. ACM Press, New York, NY, 339–346.

WISE, J., THOMAS, J., PENNOCK, K., LANTRIP, D., POTTIER, M., AND SCHUR, A. 1995. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Conference on Information Visualization*. IEEE Computer Society Press, Los Alamitos, CA, 51–58.

WITTEN, I. H., MOFFAT, A., AND BELL, T. C. 1994. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold Co., New York, NY.

WU, J. K. AND NARASIMHALU, A. D. 1994. Identifying faces using multiple retrievals. *IEEE MultiMedia 1*, 2 (Summer), 27–38.

ZAMIR, O. AND ETZIONI, O. 1998. Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '98, Melbourne, Australia, Aug. 24–28), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Chairs. ACM Press, New York, NY, 46–54.

ZAMIR, O., ETZIONI, O., MADANI, O., AND KARP, R. 1997. Fast and intuitive clustering of web documents. In *Proceedings of the ACM SIGMOD International Workshop on Data Mining and Knowledge Discovery* (SIGMOD-96, Aug.), R. Ng, Ed. ACM Press, New York, NY, 287–290.

ZHANG, T., RAMAKRISHNAN, R., AND LIVNY, M. 1996. Birch: An efficient data clustering method for large databases. In *Proceedings of the ACM-SIGMOD Conference on Management of Data* (Montreal, Canada, June). ACM, New York, NY.

ZHONG, D. AND CHANG, S.-F. 1997. Video object model and segmentation for content-based video indexing. In *Proceedings of the International Conference on Circuits and Systems*. IEEE Computer Society Press, Los Alamitos, CA.

Received: November 1998; revised: April 2000; accepted: July 2000