April 25, 2000
RT0347
Computer Science 48 pages

Research Report

Information Retrieval on the Web

Mei KOBAYASHI and Koichi TAKEDA

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Information Retrieval on the Web

Mei Kobayashi and Koichi Takeda
IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken 242-8502 Japan
mei@trl.ibm.co.jp, Kohichi_Takeda@jp.ibm.com

April 25, 2000

Abstract

In this paper we review studies on the growth of the Internet and technologies which 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. As such, it is not surprising that about 85% of Internet users surveyed claim to be using search engines and search services to find specific information of interest. The same surveys show, however, that users are not satisfied with the performance of the current generation of search engines; the slow speed of retrieval, 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 which are targeted to resolve some of the problems associated with Web-based information retrieval and speculate on future trends.

keywords: clustering, indexing, information retrieval, Internet, search engine, World Wide Web, WWW, W3.

Contents

1 Introduction

In this paper we review some notable studies on the growth of the Internet and technologies which are useful for information search and retrieval on the Web. Writing about the Web is a challenging task for at several reasons of which 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 First, its dynamic nature guarantees that at least some portions of any manuscript on the subject will be out-of-date before it reaches the intended audience, particularly URLs which are referenced. Second, a comprehensive coverage of all of the important topics is impossible, because so many new ideas are constantly being proposed and either quickly accepted into the Internet mainstream or rejected. Finally, as with any review paper, there is a strong bias in presenting topics which are 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 which describe or review concepts in depth will be given whenever possible. This having being said, we begin with references to several excellent books which cover a variety of topics in information management and retrieval.

They include: *Information Retrieval and Hypertext* [Agosti, Smeaton 1996], *Modern Information Retreival* [Baeza-Yates, Ribeiro-Neto 1999], *Text Retreival and Filtering: Analytic Models of Performance* [Losee 1998], *Natural Language Information Retreival* [Strzalkowski 1999], and *Managing Gigabytes* [Witten, Moffat, Bell 1994]. Some older, classical texts which are slightly outdated include: *Information Retreival* [Frakes, Baeza-Yates 1992], *Information Storage and Retrieval* [Korfhage 1997], *Intelligent Multmedia Information Retrieval* [Maybury 1997], *Introduction to Modern Information Retrieval* [Salton, McGill 1983], *Readings in Information Retrieval* [Sparck Jones, Willett 1997].

Additional references are special journal issues on: search engines on the Internet [Scientific American 1997], digital libraries [CACM 1998], digital libraries, representation and retrieval [IEEE PAMI 1996], the next generation graphical user interfaces (GUIs) [CACM 1994], Internet technologies [CACM 1994], [IEEE IS 1999], and knowledge discovery [CACM 1999]. Some notable survey papers are [Chakrabarti, Rajagopalan 1997], [Faloutsos, Oard 1995], [Feldman 1998], [Gudivada et al. 1997], [Leighton, Srivastava 1997] [Lawrence, Giles 1998b], [Lawrence, Giles 1999a], [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* [W3] and Association for Computing Machinery (ACM) [1] Special Interest Group on on Computer-Human Interaction [SIGCHI] and Special Interest Group on Information Retrieval [SIGIR] Conferences. A list of papers and Web pages which review and compare Web search tools are maintained at several sites, including Boutell's World Wide Web FAQ [2], Hamline University [3], Kuhn's pages [4] (in German), Maire's pages (in French) [5], Princeton University [6], U.C. Berkeley [7], and Yahoo!'s pages on search engines [8]. The historical development of information retrieval is documented in a number of sources, such as [Baeza-Yates, Ribeiro-Neto 1999], [Cleverdon 1970], [Faloutsos, 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], [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 which must be overcome in developing and refining new and existing technologies for the Web environment are discussed.

[1] *the Association for Computing Machinery* (ACM) : www.acm.org

[2] *Boutell, T.*, World Wide Web FAQ: www.boutell.com/faq/

[3] *Hamline University*: web.hamline.edu/Administration/Libraries/search/comparisons.html

[4] *Kuhn's Pages*: www.gwdg.de/h̃kuhn1/pagesuch.html#VL

[5] *Maire's pages*: www.imaginet.fr/ime/search.htm

[6] *Princeton University*: www.cs.princeton.edu/html/search.html

[7] *U.C. Berkeley*: sunsite.berkeley.edu/Help/searchdetails.html

[8] *Yahoo!*: www.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Searching_the_Web/Comparing_Search_Engines

In the concluding section we speculate on future directions of 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 on-line (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, that is, 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 which make a search or portal site highly competitive are: unusually attractive interfaces, free e-mail addresses, and free access time [Chandrasekaran 1998]. Quite often, these advatages last at most a few weeks since competitors keep track of new developments (see e.g., portalhub.com, 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, 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, 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 bi-annual WWW surveys conducted at the Graphics, Visualization, and Usability Center of Georgia Institute of Technology [9]. 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 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, Lukose 1997], [Huberman et al. 1998]. In this paper, we elaborate on

[9] GVU's user survey (available at: www.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.

some of the causes of these problems and outline some promising new approaches which are 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 [10] 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*" [11]. In addition to general advice on conducting searches on the Internet, the pages compare features of several popular search engines (i.e., *Alta Vista* [12], *HotBot* [13], *LycosPro Power Search* [14], *Excite* [15], *Yahoo!* [16], *Infoseek* [17], *Disinformation*, [18], *Northern Light* [19]) 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.

[Lidsky, Kwon 1997] is an opinionated, but informative, resource on search engines. It describes thirty six different search engines and rates them on specific details of their search capabilities. For instance, in one study, searches were 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 were evaluated in terms of power and ease of use. Variations in ratings sometimes differed substantially for a given search engine. Similarly, query tests were 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 differed 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.

[10] *AlexaInternet*: www.alexa.com/company/inthenews/webfacts.html

[11] *U.C. Berkeley*: www.lib.berkeley.edu/TeachingLib/Guides/Internet/ToolsTables.html

[12] *Alta Vista*: www.altavista.com

[13] *HotBot*: www.hotbot.com

[14] *Lycos*: www.lycos.com

[15] *Excite*: www.excite.com

[16] *Yahoo!*: www.yahoo.com

[17] *Infoseek*: www.infoseek.com

[18] *Disinformation*: www.disinfo.com

[19] *Northern Light*: www.nlsearch.com

Web Search Engine Watch [20] posts extensive data and ratings of popular search engines according to features, such as size, pages crawled per day, freshness, and depth. Other useful online sources include home pages on search engines by: Gray [21], Information Today [22], Kansas City Public Library [23], Koch [24], Northwestern University Library [25], and Notess of Search Engine Showdown [26]. Data on international use of the Web and Internet is posted at the NUA Internet Survey home page [27].

As a note of caution, we mention that 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 would 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 150 million in 2000; and Killen and Associates give the estimate 250 million in 2000. Nua's surveys give the figure 201 million worldwide in September of 1999, and more specifically by region: 1.72 million in Africa; 33.61 in the Asia/Pacific; 47.15 in Europe; 0.88 in the Middle East; 112.4 in Canada and the U.S.A.; and 5.29 in Latin America [Nua 1999]. 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

[20] *Search Engine Watch*: searchenginewatch.com/webmasters/features.html

[21] *Gray*: www.mit.people.edu/mkgray/net

[22] *Information Today*: www.infotoday.com/searcher/jun/story2.htm

[23] *Kansas City Public Library*: www.kcpl.lib.mo.us/search/srchengines.htm

[24] *Koch, T.*: www.ub2.lu.se/desire/radar/lit-about-search-services.html

[25] *Northwestern University Library*: www.library.nwu.edu/resources/internet/search/evaluate.html

[26] Notess, G., *Search Engine Showdown*: imtnet/~notes/search/index.html

[27] *NUA Internet Survey*: www.nua.ie/surveys/

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 *IEEE Internet Computing Magazine* states that exponential growth of Internet hosts was observed in separate studies by several experts [IEEE IC 1998], such as Mark Lottor of Network Wizards [28] (for a period of over ten years), Mirjan Kühne of the RIPE Network Control Center [29] (for a period of over five years in Europe), Samarada Weerahandi of Bellcore [30], and John Quarterman of Matrix Information and Directory Services [31].

The number of publicly accessible pages is also growing at an aggressive pace. [Smith 1997] estimates that in January of 1997, there were 80 million public Web pages, and that the number would subsequently double annually. [Bharat, 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 of 1997 by *Wired Digital.* Then they conducted their own experiments to estimate the size of the Web and concluded:

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

Follow up studies by [Lawrence, Giles 1999b] [32] 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. 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.

[28] *Network Wizards*: www.nw.com

[29] *RIPE Network Control Center*: www.ripe.net

[30] *Bellcore*, home page on Internet hosts: www.ripe.net

[31] *Matrix Information and Directory Services*: www.mids.org

[32] Lawrence's home page: www.neci.nec.cim/l̃awrence/papers.html

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 Wide Web Worm* [33] claims to have handled an average of 1500 queries a day in April of 1994, and *AltaVista* claims to have handled 20 million queries in November of 1997. They believe:

> "... it is likely that top search engines will handle hundreds of millions (of queries) per day by the year 2000. ..."
>
> – [Brin, 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, Schütze 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 number of valuable sites which are 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.

And 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:

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

[33] *World Wide Web Worm, home page*: wwww.cs.colorado.edu/wwww and guano.cs.colorado.edu/wwww/

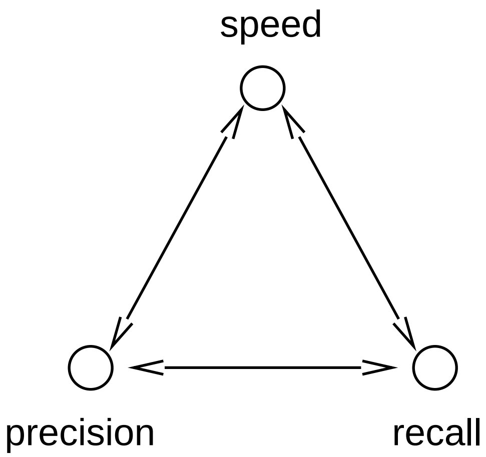

Figure 1: Three way trade-off in search engine performance: (1) speed of retrieval, (2) precision, and (3) recall.

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 which 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 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 TREC [34] search engines. In particular, they examine answers to questions, such as: *"Can link information result in better rankings ?"* and *"Do longer queries result in better answers ?"*.

[34] U.S. National Institute of Standards and Technology (NIST) Text Retrieval Conferences (TREC): trec.nist.gov

2 Tools for Web-Based Retrieval and Ranking

Classical retrieval and ranking algorithms which were 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 which 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, which set apart classical and Web-based retrieval problems. If we consider the set of all Web pages as a gigantic database, this set is very different from a classical database with elements which can be organized, stored, and indexed in a manner which 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 and 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] 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, Ribeiro-Neto 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, 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 different 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 it be 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 of 1997, the lower bound on indexable Web pages is 320 million, and a given individual search engine will have indexed between 3% to 34% of the possible total [Lawrence, 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 Feb. 2–28, 1999, involving the top 11 search engines (the above six plus: *Snap* [35], *Microsoft* [36], *Google* [37], *Yahoo!* and *Euroseek* [38]) 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, Giles 1999b].

Bharat and Broder estimated in November of 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) [39] reports weekly on indexing coverage and quality of indexing by a few, select search engines, which claim to index "at least one fifth of the Web". Other studies on estimating the extent of Web pages which have been indexed by popular search engines include [Baldonado, Winograd 1997], [Hernández 1996], [Hernández, Stolfo 1995], [Hylton 1996], [Monge, Elkan], [Selberg, Etzioni 1997], [Silberschatz et al. 1995].

[35] *Snap*: www.snap.com

[36] *Microsoft*: www.msn.com

[37] *Google*: www.google.com

[38] *Euroseek*: www.euroseek.com

[39] *Melee's Indexing Coverage Analysis (MICA) Report*: www.melee.com/mica/index.html

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 [40]; the contents usually contain non-textual multimedia data; and updates to the pages are made at different rates. [Brake 1997] estimates that the average page of text remains unchanged on the Web for about 75 days, and [Kahle 1997] estimates that 40% of the Web changes every month. Multiple copies of identical or near identical pages are abundant, for example, FAQs [41] postings, mirror sites, old and updated versions of news and newspaper sites. [Broder et al. 1997] and [Shivakumar, Garcı́a-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, Winograd 1997], [Hernández 1996], [Hernández, Stolfo 1995], [Hylton 1996], [Monge, Elkan], [Selberg, Etzioni 1997], [Silberschatz et al. 1995].

[Henzinger et al. 1999] has 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 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 a Web pages according to its *indegree* (an evaluation measure which is based on how many other pages point to the Web page under consideration [Carriere, Kazman 1997]) and *PageRank* (an evaluation measure which is based on how many other pages point to the Web page under consideration as well as the value of the pages which point to it [Brin, Page 1998], [Cho, Garcia-Molina, Page 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 to indexing is the use of *metadata*, i.e., summaries of Web page content or sites which are placed in the page for the purpose of aiding automatic indexers.

[40] 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.

[41] 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: www.cis.ohio-state.edu/hypertext/faq/usenet/FAQ-List.html and www.faq.org .

2.1.1 Classical Methods

Manual indexing is currently used by several commercial, Web-based search engines, e.g., *Galaxy* (TradeWave, formerly EINet) [42], *GNN: Whole Internet Catalog* [43], *Infomine* [44], *KidsClick!* [45], *LookSmart* [46], *Subject Tree* [47], *Web Developer's Virtual Library* [48], *World-Wide Web Virtual Library Series Subject Catalog* [49], and *Yahoo!*. And the practice is unlikely to continue to be as successful over the next few years, however, 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 two different professional indexers; as little as 20% of the terms to be indexed may be handled in the same manner by different individuals (as noted in [Korfhage 1997], p. 107), and there is noticeable inconsistency, even by a given individual [Borko 1979], [Cooper 1969], [Jacoby, Slamecka 1962], [Macskassy et al. 1998], [Preschel 1972], [Salton 1969].

Though not perfect, compared to most automatic indexers, human indexing is currently the most accurate since 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], [Sparck Jones, Willett 1997], [van Rijsbergen 1977], Chapter 3 in [Witten, Moffat, Bell 1994]. Technological advances are expected to narrow the gap in indexing quality between human and machine generated indexes. In the future, human indexing will inly 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

Recently, scientists are investigating the use of *intelligent agents* for performing specific tasks, such as indexing on the Web [50] [AI Magazine 1997], [Baeza-Yates, 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, crawlers, Web robots* 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-

[42] *Galaxy*: galaxy.einet.net

[43] *Whole Internet Catalog*: www-e1c.gnn.com/gnn/wic/index.html

[44] *Infomine*: lib-www.ucr.edu

[45] *KidsClick!*: sunsite.berkeley.edu/KidsClick!/

[46] *LookSmart*: www.looksmart.com

[47] *Subject Tree*: www.bubl.bath.ac.uk/BUBL/cattree.html

[48] *Web Developer's Virtual Library*: www.stars.com

[49] *World-Wide Web Virtual Library Series Subject Catalog*: www.w3.org/hypertext/DataSources/bySubject/Overview.html

[50] *Web robots FAQ*: info.webcrawler.com/mak/projects/robots/faq.html

combination with other technologies. Examples are: *Aliweb* [51]; *AltaVista*; *Excite*; *Harvest* [52]; *HotBot*; *Infoseek*; *Lycos*; *Magellan* [53]; *MerzScope* [54]; *Northern Light*; *Smart Spider* [55]; *Webcrawler* [56]; 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 are: *Argos* [57] for web sites on the ancient and medieval worlds *CACTVS Chemistry Spider* [58] for chemical databases; *MathSearch* [59] for English mathematics and statistics documents; *NEC-MeshExplorer*, [60] for the NETPLAZA search service owned by NEC Corporation; and *Social Science Information Gateway* [61] *(SOSIG)* for resources in the social sciences. Crawlers which index documents in limited environments include: *LookSmart* [62] for a 300,000 site database of rated and reviewed sites; *Robbie the Robot* which is funded by DARPA of the United States government for education and training purposes; and *UCSD Crawl* [63] for UCSD pages. More extensive lists of intelligent agents are available on the data base of robots on *The Web Robots Page* [64] and Washington State University's robot pages [65].

To date, three major problems have been 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* [66] and Bloomberg [67] and every few hours by many newspaper sites [Carl 1995], [Koster 1995]. Some Web sites deliberately keep out spiders, for example: the *New York Times* [68] requires users to pay and fill out a registration form; *CNN* used to exclude search spiders to prevent distortion of data on number of users who visit the site; and the online catalogue of the *British Library* [69] 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 which will protect their sites from new types of agents which intentionally or unintentionally may cause mayhem

[51] *Aliweb*: www.nexor.co.uk/public/aliweb/aliweb.html

[52] *Harvest*: harvest.transarc.com

[53] *Magellan*: www.magellan.com

[54] *MerzScope*: www.merzcom.com

[55] *Smart Spider*: www.engsoftware.com

[56] *Webcrawler*: webcrawler.com/

[57] *Argos*: argos.evansville.edu

[58] *CACTVS Chemistry Spider*: schiele.organik.uni-erlangen.de/cactvs/spider.html

[59] *MathSearch*: www.maths.usyd.edu.au:8000/MathSearch.html

[60] *NEC-MeshExplorer*: netplaza.biglobe.or.jp/keyword.html

[61] *SOSIG*: scout.cs.wisc.edu/scout/mirrors/sosig

[62] *LookSmart*: www.looksmart.com/

[63] *UCSD Crawl*: www.mib.org/~ucsdcrawl

[64] *Web Robots Page*: info.webcrawler.com/mak/projects/robots/active/html/type.html

[65] *WSU pages by by Felt and Scales*: www.wsulibs.wsu.edu/general/robots.htm

[66] *CNN*: www.cnn.com

[67] *Bloomberg*: www.bloomberg.com

[68] *New York Times*: www.nytimes.com

[69] *British Library*: portico.bluk

in their computer systems.

As a working compromise, Kostner has proposed a *robots exclusion standard* [70], which advocates blocking certain types of searches to relieve overload problems. He has also proposed guidelines for *robot design* [71]. 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], [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 which 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* [72] has compiled a list of resources on *information and standardization proposals for metadata* [73]. A number of metadata standards have been proposed for Web pages. Among them, two well-publicized, solid efforts are: the *Dublin Core Metadata standard* [74] uand the *Warwick framework* [Lagoze 1996]. The Dublin Core is a 15-element metadata element set proposed for the purpose of facilitating 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, Ribeiro-Neto 1999], [Marchionini 1995]. If the general public adopts and increases use of a simple metadata standard

[70] Koster, M., "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

[71] Koster, M., "Guidelines for robot writers" (1993):
info.webcrawler.com/mak/projects/robots/guidelines.html

[72] *World Wide Web (W3) Consortium*: www.w3.org

[73] *W3 metadata page*: www.w3.org/Metadata

[74] *Dublin Core Metadata*, home page: purl.oclc.org/metadata/dublin_core

(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 a 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 which can occur 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)* [75] be used as the metadata coding scheme for Web documents. An interesting associated development is IBM's *XCentral* [76], 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 in order to increase legitimate traffic to a site. Unfortunately, not all Web page authors are fair players. Many unfair players maintain sites which 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.

Related to, but separate from the unethical or deceptive use of metadata is *spamming*, i.e., excessive, repeated use of key words or "hidden" text which are purposely inserted into a Web page to promote retrieval by search engines. Spamming is a new phenomenon which appeared with the introduction of search engines, automatic indexers, and filters on the Web [Flynn 1996], [Liberatore 1998]. 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, ego-centric Web page authors wanting attention, and technically well-versed, but off-balanced individuals who have the same sort of warped mentality as inventors of computer viruses. A famous example of hidden text spamming was the embedding of words in a black background by the *Heaven's Gate Cult* [77], a technique which has come to be known as *font color spamming* [Liberatore 1998]. We note that the term *spamming* has a broader meaning, related to the receiving of excessive amount of email or information. An excellent, broad overview of the subject is given in [Cranor, LaMacchia 1998]. In the context we are considering, the specialized terms *spam-indexing*, *spam-dexing*, or *keyword spamming* are more precise.

[75] *W3 Consortium RDF homepage:* www.w3.org/RDF

[76] *IBM's XCentral homepage:* www.ibm.com/developer/xml

[77] Although the cult no longer exists, the Heaven's Gate Cult home page is archived at the sunspot.net site: www.sunspot.net/news/special/heavensgatesite

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, Hasida 1998], [Nagao et al. 1999]. Three examples of annotations which are most common 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 non-textual 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 in indexing and retrieval, this kind of annotation may be helpful for evaluating documents.

Despite the promise that use of metadata and annotation may facilitate fast and accurate search and retrieval, a recent study for the period Feb. 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, Giles 1999b]. Unless a new trend towards the use of metadata and annotations develops, then their usefulness in information retrieval may be limited to very large, closed data owned by large corporations, public institutions and governments which choose to use them.

2.2 Clustering

The grouping together of similar documents to expedite information retrieval is known as *clustering* [Anick, Vaithyanathan 1997], [Rasmussen 1992], [Sneath, 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 which 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 *non-hierarchical.* Hierarchical methods show greater promise for enhancing Internet search and retrieval systems. Although details of clustering algorithms used by major search engines is not publicly available, some general approaches are known. For instance, the Web search engine *AltaVista* of Digital Equipment Corporation (DEC) is based on clustering. [Anick, Vaithyanathan 1997] explores 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 which 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 in at least two orders of magnitude faster than methods which produce comparable clusters. A third method, called *suffix tree clustering* is detailed in [Zamir, Etzioni 1998].

[Modha, Spangler 1999] 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 out-links, 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 have been designed specifically 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 1998], [Dhillon 1999] [Ester et al. 1995a], [Ester et al. 1995b], [Ester et al. 1995c], [Fisher 1995], [Guha et al. 1998], [Ng, Han 1994], [Zhang et al. 1996]. For very large databases, apropriate parallel algorithms can speed up computations [Omiecinski, 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, Baeza-Yates 1992]. Specific examples of some methods designed specifically for facilitating Web-based information retrieval are: evaluation of the significance, reliability and topics covered in a set of Web pages based on analysis of the hyper-link 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 non-textual data (e.g., hand drawn sketches, textures and colors, speech) and better user interfaces to enable users to visually manipulate retrieved information [Card, MacKinlay, Shneiderman 1999], [Hearst 1997], [Maybury, 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 Meta-search Navigators

A very simple tool developed to exploit the best features of many search engines is the meta search navigator. These navigators allow simultaneous search of a set of other navigators. Two of the most extensive ones are *Search.com* [78], which can utilize the power of over 250 search engines and *INFOMINE* [79] which utilizes over 90. Advanced meta-search navigators have a single input interface which sends queries to all (or only user selected search engines), eliminates duplicates, then combines and ranks returned results from the different search engines. Some fairly simple examples available on the Web are: *2ask* [80], *ALL-IN-ONE* [81], *EZ-Find at The River* [82], *IBM InfoMarket Service* [83], *Inference Find* [84], *Internet Sleuth* [85], *MetaCrawler* [86], and *SavvySearch* [Howe, Dreilinger 1997] [87].

2.3.2 Web-Based Information Outlining/Visualization

Visualization tools designed specifically for helping users understand Websites (e.g., their directory structures, types of information available) are being developed by many private and public research centers [Nielsen 1999]. Overviews of some of these tools are given in [Ahlberg, Shneiderman 1994], [Beaudoin, Parent, Vroomen 1996], [Bederson, Hollan 1994], [Gloor, Dynes 1998], [Lamping, Rao, Pirolli 1995], [Liechti, Sifer, Ichikawa 1998], [Maarek et al. 1997], [Munzner, Burchard 1995], [Robertson, MacKinlay, Card 1991], [Tetranet 1998]. Below we present some examples of interfaces which were designed for facilitating general information retrieval systems, then we present some which 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 which: allow real estate brokers and their clients to locate homes based on price, number of bedrooms, distance from work, etc. [Williamson, Shneiderman 1992]; locate geographical regions which have cancer rates above the national average [Plaisant 1994]; allow dynamic queryiing of a chemistry table

[78] *Search.com*: www.search.com/

[79] *INFOMINE*: lib-www.ucr.edu/enbinfo.html

[80] *2ask*: web.gazeta.pl/ miki/search/2ask-anim.html

[81] *ALL-IN-ONE*: www.albany.net/allinone/

[82] *EZ-Find at The River*: www.theriver.com/TheRiver/Explore/ezfind.html

[83] *IBM InfoMarket Service*: www.infomkt.ibm.com/

[84] *Inference Find*: www.inference.com/infind/

[85] *Internet Sleuth*: www.intbc.com/sleuth

[86] *MetaCrawler*: metacrawler.cs.washington.edu:8080/

[87] *SavvySearch*: www.cs.colostate.edu/d̃reiling/smartform.html
and guaraldi.cs.colostate.edu:2000/

[Ahlberg, Shneiderman 1997]; with an interface to enable users to explore UNIX directories through dynamic queries [Liao et al. 1992]. Features of the systems are: visual presentation of query components; visual presentation of results; rapid, incremental and reversible actions; selection by pointing (not typing); and immediate and continuous feedback. Most graphics hardware systems in the mid-1990's were still too weak to provide adequate real-time interaction, however, faster algorithms and advances in hardware should increase the speed up of the system 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 a queries by allowing users to critique retrieved results with labels, such as "*require*" and "*prohibit*". Williams claims that the 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], [Hearst, 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 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 about the document, and the information is encoded as a number, the magnitude of which is represented according to a grayscale.

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

[Baldonado 1997], [Baldonado, 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 a 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 together results whose URLs refer to the same site; (2) bundling together 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 URL, and also allows users to select from several criteria 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 has been investigated extensively by many other scientists, e.g., [Hernández 1996], [Hernández, Stolfo 1995], [Hylton 1996], [Monge, Elkan], [Silberschatz et al. 1995].

[Card, Robertson, York 1996] developed two 3D virtual interface tools, *WebBook*, and *Web-*

Forager, 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, such as query refinement and understanding the contents of a databse from different perspectives (each query represents a different user perspective). [Manber, Smith, Gopal 1997] developed *WebGlimpse*, a tool for simultaneous searching and browsing of Web pages, which is based on the *Glimpse* search engine.

[Morohashi et al. 1995], [Takeda, Nomiyama 1997] developed a system, which uses new technologies to organize and display, in an easy 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 the multiple viewers facilitates the reduction of candidate results, profiling of the information, and discovering of new facts. And [Sakairi 1999] developed a site map for visualizing a Web site's structure and keywords

2.3.3 Acoustical Interfaces

Web-based IR is contributing to the acceleration of studies on and development of more user friendly, non-visual, input-output interfaces. Some examples of research projects are given in a special journal issue devoted to the topic "*the Next Generation Graphics User Interfaces (GUIs)*" [CACM 1993]. The article discusses how some users have indicated a 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) [Business Week 1998].

One answer 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 [88]. The basic idea of the method is to identify different languages (e.g., English, Japanese) and different text types (e.g., title and section headers, regular text, hot buttons), then assign 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, Asakawa 1998].

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

A third acoustic-based approach for Web browsing is being investigated by [Mereu, Kazman 1996]. They examined how sound environments can be used for navigation and found that sighted users prefer the use of 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 use in more general systems (i.e., not necessarily for the visually impaired) which require an audio/speech based interface.

[88] *IBM Homepage on Systems for the Disabled*: www.trl.ibm.co.jp/projects/s7260/sysd_e.htm

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 will give references to some classical techniques which can be modified for use by Web search engines [Baeza-Yates, Ribeiro-Neto 1999], [Berry, Browne 1999], [Frakes, Baeza-Yates 1992]. Techniques developed specifically for the Web will also be 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 of or vector space models [Baeza-Yates, 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 which can help in distinguishing 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 [89]. Ranking of 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 dimension 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 which 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

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

largest singular values σ_i ; $i = 1, 2, ..., k$, 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 [90].

Processing the query takes place in two steps: projection, followed by matching. In the projection step, input queries are mapped to pseudo-documents 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

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

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

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

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

A simple means of measuring the quality of a Web page proposed by [Carriere, Kazman 1997] is to count the number of pages which have pointers to the page is used in the *WebQuery* system and the *Rankdex* search engine [91]. Another search engines which use link infomation is *Google*, which currently indexes about 85 million Web pages. Its rankings are based, in part, on the number of other pages which have pointers to the page. This policy seems to favor slightly educational and government sites over commercial sites. In November of 1999, Northern Light introduced a new ranking system, which is also based, in part, on link data [92].

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

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

[91] *Rankdex*: rankdex.gari.com

[92] Search Engine Briefs Nov. 1999: www.searchenginewatch.com/sereport/99/11-briefs.html

[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 which have 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 which are likely to make an impact on Web use in some form or variation in the future. Knowledge management [IEEE IS 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 IS 1996]. Problems which can be tackled by these agents include: finding and filtering information; customizing information; and automating completion of simple tasks [Gilbert 1997]. These agents "gather information or perform some other service without (the user's) immediate presence and on some regular schedule" [93]. The *BotSpot* Web site [94] summarizes and points to some historical information as well as current work on intelligent agents. The *Procedings of the Association for Computing Machinery* [95] *(ACM), Conferences on Information and Knowledge Management* (CIKM) and the *American Association for Artificial Intelligence Workshops* [96] are valuable information sources. The *Proceedings of the Practical Applications of Intelligent Agents and Multi-Agents* (PAAM) conference series [97] gives a nice overview of application areas. The home page of the *IBM Intelligent Agent Center of Competence* (IACC) [98] describes some of the company's commercial agent products and technologies for the Web.

One interesting area in intelligent Web robot research is adaptive Web services. Examples of services include: *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, Weld 1995], [Perkowitz, Etzioni 1999]; and *Adaptive Web Page Recommendation Service* [Balabanović 1997], [Balabanović, Shoham], [Balabanović et al. 1995]. Discussion and ratings of some of these and other robots are available at several Web sites, e.g., [Felt, Scales], [Mitchell].

Some scientists have studied prototype *metasearchers*, i.e., services which 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, Giles 1998a], [Selberg, Etzioni 1995], [Selberg, Etzioni 1995]. Some of the more well known meta search engines include: *MetaCrawler,*

[93] *whatis?com home page*: www.whatis.com/intellig.htm

[94] *BotSpot home page*: botspot.com

[95] for URL see section 1.5

[96] *American Association for Artificial Intelligence home page*: www.aaai.org

[97] *PAAM96*: www.demon.co.uk/ar/PAAM96 , and *PAAM97*: www.demon.co.uk/ar/PAAM97

[98] *IBM-IACC*: www.networking.ibm.com/iag/iaghome.html

SavvySearch and *InfoSeek Express.* Metasearchers work in three main steps after a query is issued: 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, identification of duplicate (and near duplicate) pages [Lawrence, 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 on-line 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, the founder of *Amazon.com* for the 1999 *Person of the Year.* Exponential growth is predicted in on-line shopping. Charts which 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 [Nua 1999].

Robots to help consumers shop, or *shopbots*, have become commonplace in e-commerce sites and general purpose Web-portals. Shopbot technology has 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 of Frictionless Commerce (www.frictionless.com) footnoteFrictionless’s bot engine is used by some famous portals, including Lycos mentioned that his company’s technology will be used in a retailer bot which 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. 1998], [Kephart, Hanson, Sairamesh 1998].

Another successful technological off-shoot of the Internet shopping business are auction sites [Cohen 1999], [Ferguson 1999]. Two of the more famous general on-line aution sites are *priceline.com* [99] and *eBay* [100]. Priceline.com pioneered and patented their business concept of on-line bidding [Walker, Sparico, Case 1999]. Patents related to that of priceline.com include ones owned by ADT Automotive, Inc. [Berent et al. 1998], Walker Asset Management [Walker et al. 1998], and two individuals [Barzilai, Davidson 2000].

3.3 Multimedia Retrieval

IR from multimedia databases is a multidisciplinary area of research, which includes topics from a very diverse range, such as analysis of text, image and video, speech, and non-speech audio; graphics; animations; artificial intelligence; human-computer interaction; and multimedia computing [Faloutsos 1996], [Faloutsos, Lin 1995], [Maybury 1997], [Schäuble 1999]. Recently, several commercial systems which integrate search capabilities from multiple databases containing heterogeneous, multimedia data have become available. Examples include: PLS [101], Lexis-Nexis [102], DIALOG [103], and Verity [104]. In this section we point to some recent developments in the field; The discussion is by no means comprehensive.

One of the more established fields of research involving multimedia databases is query and retrieval of images [ICIP], [ICASSP], [IFIP 1989, 1992]. So much work on the topic has been conducted by many, that a comprehensive review is beyond the scope and purposes of this paper. Some selected works in this area are: search and retrieval from large image archives [Castelli et al. 1998]; pictorial queries by image similarity [Soffer, Samet 1999]; image queries using Gabor wavelets features [Manjunath, Ma 1996]; fast, multiresolution image queries using Haar wavelet transform coefficients [Jacobs et al. 1995]; acquisition, storage, indexing and retrieval of map images [Samet, Soffer 1996]; real-time fingerprint matching from a very large database [Ratha et al. 1996]; querying and retrieval using partially decoded JPEG data and keys [Schneier, Abdel-Mottaleb 1996]; and retrieval of faces from a database [Bach et al. 1993], [Wu, Narasimhalu 1994].

Finding documents on the Web which have images of interest is a much more sophisticated problem. Two well-known portals which have a search interface for a database of images are: *Yahoo! Image Surfer* [105] and *Alta Vista PhotoFinder* [106]. Like Yahoo!'s text-based search engine, the Image Surfer home pages are organized into categories. For a text-based query, a

[99] *priceline.com*: www.priceline.com

[100] *eBay*: www.ebay.com

[101] *PLS homepage*: www.pls.com

[102] *Lexis-Nexis homepage*: www.lexis-nexis.com

[103] *DIALOG homepage*: www.dialog.com

[104] *Verity homepage*: www.verity.com

[105] *Yahoo! Image Surfer*: isurf.yahoo.com

[106] *Alta Vista PhotoFinder*: image.altavista.com

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 movies 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-base 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 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 which have 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; the most of the thumbnails were relevant for the first several pages (each page contained 10 thumbnails).

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

A research area closely related to retrieval of still images from a very large image database is query and retrieval of images in a video frame or frames [Bolle 1998]. We mention just a few selected projects to illustrate the potentially wide scope of applications. Some examples are: content-based video indexing retrieval [Smoliar 1994]; the *Query-by-Image-Content (QBIC)* system, which helps users find still images in large image and video databases based on color, shape, texture, and sketches [Flickner et al. 1997], [Niblack et al. 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, Chang 1997], [Smith, Chang 1996b]; compressed domain image manipulation and feature extraction for compressed domain image and video indexing and searching [Chang 1995], [Zhong, Chang 1997]; a method for extracting visual events in a relatively long videosuing objects (rather than keywords), with specific applications to videos of 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 perform identity checks on fish [Sze, Liao, Lu 2000]; and searching for images and videos on the Web [Smith, Chang 1996a].

Multilingual communication on the Web [Miyahara et al. 2000] and cross-language document retrieval is a timely research topic which is being investigated by many institutions [Ballesteros, Croft 1998], [Eichmann, Ruiz, Srinivasan 1998], [Pirkola 1998]. An introduction to the subject is given in [Oard 1997c] and some surveys are [CLIR], [Oard 1997b], [Oard, Dorr 1996]. Several search engines now feature multilingual search. For example, *Open*

Text Web Index [107], searches in 4 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 [108]. A typical example, which has a trial version for downloading is product called "*Honyaku no Oosama*" [109] or Internet King of Translation [Watanabe, Takeda 1998].

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

3.4 Conclusions

Potentially lucrative applications 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 number of hosts are not sustainable so that research and business opportunities on the subject 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 with slow access times and an August 1998 survey by Alexa Internet which states 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, these pessimists maintain that effective means of managing the highly uneven concentration of information packets on the Internet are not immediately available nor will 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, Boroumand 2000] maintain that flat rate Internet retailing 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 continuation of the current solution, i.e., overprovisioning of bandwidth.

Acknowledgments

The authors acknowledge helpful conversations with Stuart McDonald of alphaWorks and our colleagues in IBM Research. Our manuscript has benefitted greatly from the extensive and

[107] *Open Text Web Index*: index.opentext.net

[108] *homepage on E-to-J SW* (in Japanese): www.bekkoame.ne.jp/ oto3/

[109] *Honyaku no Oosama*: www.ibm.co.jp/software/internet/king/index.html

well-documented list of suggestions and corrections from the reviewers of the first draft. We appreciate their generosity, patience and thoughtfulness.

References

[Agosti, Smeaton 1996] Agosti, M., Smeaton, A., *Information Retrieval and Hypertext*, Kluwer, Boston (1996).

[Agrawal et al. 1998] Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., "Automatic subspace clustering for high dimensional data for data mining", *Proc. ACM SIGMOD Conference on Management of Data*, ACM Press, NY (1998).

[Ahlberg, Shneiderman 1994] Ahlberg, C., Shneiderman, B., "Visual information seeking: tight coupling of dynamic query filters with starfield displays", *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, Apr. 24–28, Boston, MA (1994), 313–317.

[Ahlberg, Shneiderman 1997] Ahlberg, C., Shneiderman, B., "Alphaslider: a rapid and compact selector, *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems* (1994).

[AI Magazine 1997] *AI Magazine*, special issue on intelligent systems on the Internet (Summer 1997).

[American Heritage Dictionary] *American Heritage Dictionary*, Houghton Mifflin Co., Boston, MA (1976).

[Anderberg 1973] Anderberg, M., *Cluster Analysis for Applications*, Academic Press, NY (1973).

[Anick, Vaithyanathan 1997] Anick, P., Vaithyanathan, S., "Exploiting clustering and phrases for context-based information retrieval", *Proc. ACM Special Interest Group on Information Retreival (SIGIR)*, ACM Press, NY (1997), 314–323.

[Asakawa 1996] Asakawa, C., "Enabling the visually disabled to use the WWW in a GUI environment", *Technical Report of the IEICE*, HC96-29, Tokyo, Japan (Sept. 1996), 39–44.

[Bach et al. 1993] Bach, J., Paul, S., Jain, R., "A visual information management system for the interactive retrieval of faces" *IEEE Trans. Knowledge and Data Engineering*, 5, 4 (Aug. 1993), 619–628.

[Baeza-Yates, Ribeiro-Neto 1999] Baeza-Yates, R., "Introduction to data structures and algorithms related to information retrieval", in Baeza-Yates, R., Ribeiro-Neto, B. (eds.), *Modern Information Retreival*, ACM Press, New York (1999) 13–27.

[Baeza-Yates, Ribeiro-Neto 1999] Baeza-Yates, R., Ribeiro-Neto, B. (eds.), *Modern Information Retreival*, ACM Press, New York (1999).

[Balabanović 1997] Balabanović, M. "An adaptive Web page recommendation service", *Stanford Univ. Digital Libraries Project Working Paper* DISL-WP 1996-0041, also available in *Proc. First Int'l. Conference on Autonomous Agents*, Martina del Rey, CA (Feb. 1997).

[Balabanović, Shoham] Balabanović, M., Shoham, Y., "Learning information retrieval agents: experiments with automated Web browsing", preprint, Stanford Univ., Dept. Computer Science.

[Balabanović et al. 1995] Balabanović, M., Shoham, Y., Yun, T., "An adaptive agent for automated Web browsing", *Stanford Univ. Digital Libraries Project, working paper* DISL-WP 1995-0023 (1995).

[Baldonado 1997] Baldonado, M., *An interactive, structure-mediated approach to exploring information in a heterogeneous, distributed environment*, Ph.D. Thesis, Dept. Computer Science, Stanford Univ. (Dec. 1997).

[Baldonado, Winograd 1997] Baldonado, M., Winograd, T., "Sensemaker: an information-exploration interface supporting the contextual evolution of a user's interests", *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, Atlanta, Georgia (March 1997), 11–18.

[Ballesteros, Croft 1998] Ballesteros, L., Croft, B., "Resolving ambiguity for cross-language retrieval", *Proc. ACM Special Interest Group on Information Retreival (SIGIR)*, ACM Press, NY (1998), 64–71.

[Barzilai, Davidson 2000] "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, filed July 1, 1997, issued Jan. 4, 2000.

[Beaudoin, Parent, Vroomen 1996] Beaudoin, L., Parent, M.-A., Vroomen, L., "Cheops: a compact explorer for complex hierarchies", *Proc. IEEE Visualization* (Oct./Nov. 1996) 87–92.

[Bederson, Hollan 1994] Bederson, B., Hollan, J., "Pad++: a zooming graphical interface for exploring alternate interface physics", *Proc. ACM Symposium on User Interface Software and Technology* (Nov. 1994), 17–26.

[Berent et al. 1998] Berent, T., Hurst, D., Patton, T., Tabernik, T., Warpool, J., Reig, D., Whittle, W., "Electronic on-line motor vehicle auction and information system", U.S. Patent no. 5774873, filed March 29, 1996, issued June 30, 1998.

[Berners-Lee et al. 1994] Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H., Secret, A., "The World-Wide Web", *Communications of the ACM*, 37, 8 (Aug. 1994), 76–82.

[Berry, Browne 1999] Berry, M., Browne, M., *Understanding Search Engines*, SIAM, Philadelphia (1999).

[Berry et al. 1995a] Berry, M., Dumais, S., O'Brien, G., "Using linear algebra for intelligent information retrieval", *SIAM Review*, 37, 4 (Dec. 1995), 573–595.

[Bharat, Broder 1998] Bharat, K., Broder, A., "A technique for measuring the relative size and overlap of public Web search engines", *Proc. 7th World Wide Web Conference* (1998): www7.conf.au/programme/fullpapers/1937/com1937.htm

[Bolle 1998] Bolle, R., Yeo., B.-L., Yeung, M., "Video query: research directions", *IBM Journal of Research and Development*, 42 (2) (March 1998), 233–251.

[Borko 1979] Borko, H., "Inter-indexer consistency", *Cranfield Conference* (1979).

[Botafogo et al. 1992] Botafogo, R., Rivlin, E., Shneiderman, B., "Structural analysis of hypertext: identifying hierarchies and useful metrics", *ACM Trans. on Information Systems*, 10 (1992), 1412–180.

[Brake 1997] Brake, D., "Lost in cyberspace", *New Scientist Magazine* (June 28, 1997): www.newscientist.com/keysites/networld/lost.html

[Brin, Page 1998] Brin, S., Page, L., "The anatomy of a large-scale hypertextual Web search engine", preprint, Dept. of Computer Science, Stanford Univ. (1998).

[Broder et al. 1997] Broder, A. et al., "Syntatactic Clustering of the Web", *Proc. Sixth Int'l WWW Conference*, Santa Slara, CA (Apr. 1997), 391–404.

[Business Week 1998] *Business Week*, Special report on speech technologies (Feb. 23, 1998).

[Card, MacKinlay, Shneiderman 1999] Card, S., Mackinlay, J., Shneiderman, B. (eds.), *Readings in Information Visualization*, Morgan Kaufmann, San Francisco, CA (1999).

[Card, Robertson, York 1996] Card, S., Robertson, G., York, W., "The WebBook and WebForager: an information workspace for the World-Wide Web, *proc. ACM Conference on Human Factors in Computing Systems*, ACM Press, New York (1996) 111-117.

[Carl 1995] Carl, J. "Protocol gives sites way to keep out the 'bots'", *Web Week*, 1, 7 (Nov. 1995): info.webcrawler.com/mak/projects/robots/threat-or-treat.html

[Carriere, Kazman 1997] Carriere, J., Kazman, R., "WebQuery: searching and visualizing the Web through connectivity", *Proc. of the Sixth Int'l. World Wide Web Conference* (1997): www.w3.org/Conferences/Overview-WWW.html

[Castelli et al. 1998] Castelli, V., Bergman, L., Kontoyiannins, I., Li, C.-S., Robinson, J., Turek, J., "Progressive search and retrieval in large image archives", *IBM Journal of Research and Development*, 42 (2), (March 1998), 253–268.

[Cathro 1997] Cathro, W., "Matching discovery and recovery", *the Standards Australia Seminar* (Aug. 1997): www.nla.gov.au/staffpaper/cathro3.html

[Chakrabarti, Rajagopalan 1997] Chakrabarti, S., Rajagopalan, S., "Survey of information retrieval research and products", last update (Apr. 24, 1997):
w3.almaden.ibm.com/s̃oumen/ir.html

[Chakrabarti et al. 1998a] Chakrabarti, S., Dom, B., Gibson, D., Kumar, S., Raghavan, P., Rajagopalan, S., Tomkins, A., "Experiments in topic distillation", *ACM-SIGIR Post Conference Workshop on Hypertext Information Retrieval for the Web*, Melbourne, Australia (April 1988).

[Chakrabarti et al. 1998b] Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., Kleinberg, J., "Automatic resource compilation by analyzing hyperlink structure and associated text", *Proc. of the Seventh Int'l. World Wide Web Conference* (1998):
www7.conf.au/programme/fullpapers/1898/com1898.htm

[Chalmers, Chitson 1992] Chalmers, M., Chitson, P., "Bead: exploration in information visualization", *Proc. ACM Special Interest Group on Information Retreival (SIGIR)*, ACM Press, NY (1992) 330–337.

[Chandrasekaran 1998] Chandrasekaran, R., "'Portals' offer one-stop surfing on the Net", *Int'l Herald Tribune* (Oct. 12, 1998) 19 & 21.

[Chang 1995] Chang, S.-F., "Compressed domain techniques for image/video indexing and manipulation", *Proc. ICIP '95*, Vol. 1, IEEE Press, Piscataway, NJ (1995) 314–317.

[Cho, Garcia-Molina, Page 1998] Cho, J., Garcia-Molina, H., Page, L., "Efficient crawling through URL ordering", *Computer Networks* (special issue: Proc. 7th World Wide Web Conference), 30 (1998), 161–172.

[Clark 2000] Clark, D., "Shopbots become agents for business change", *IEEE Computer*, (Feb. 2000), 18–21.

[Cleverdon 1970] Cleverdon, C., "Progress in documentation", *Journal of Documentation*, 26, (1970), 55-67.

[Cohen 1999] Cohen, A., "The attic of *e*", *Time* (Dec. 27, 1999 – Jan. 3, 2000).

[CACM 1993] *Communications of the ACM*, special issue on the next generation GUIs (Apr. 1993).

[CACM 1994] *Communications of the ACM*, special issue on Internet technology (Aug. 1994).

[CACM 1998] *Communications of the ACM*, special issues on digital libraries (Apr. 1995 and Apr. 1998).

[CACM 1999] *Communications of the ACM*, special issues on knowledge discovery (Nov. 1999).

[Cooper 1969] Cooper, W., "Is interindexer consistency a hobgoblin?", *Americam Documentation*, 20, 3 (1969), 268–278.

[Cranor, LaMacchia 1998] Cranor, L., LaMacchia, B., "Spam!", *Communications of the ACM*, 41, 8 (Aug. 1998), 74–83.

[Crestani et al. 1998] Crestani, F., Lalmas, M., van Rijsbergen, C., Campbell, I., "'Is this document relevant ? ... probably': a survey of probablistic models in information retrieval", *ACM Computing Surveys*, 30, 4 (Dec. 1998).

[CLIR] Cross-Language Information Retreival Project, Univ. of Maryland, College Park:
www.clis.umd.edu/dlrg resource page: www.clis.umd.edu/dlrg/clir/papers.html
bibliography of papers on the subject: www.clis.umd.edu/dlrg/clir/bibtex.txt

[Cunningham 1997] Cunningham, M., "Brewster's millions", *Irish Times* (Jan. 27, 1997):
www.irish-times.com/irish-times/paper/1997/0127/cmp1.html

[Cutting et al. 1993] Cutting, D., Karger, D., Pedersen, J., "Constant interaction time Scatter/Gather browsing of very large document collections", *Proc. ACM Special Interest Group on Information Retreival (SIGIR)*, ACM Press, NY (1993).

[Deerwester et al. 1990] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R., "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, 41, 6 (1990), 391–407.

[Demmel 1997] Demmel, J., *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA (1997).

[Dhillon 1998] Dhillon, I., Modha, D., "A data-clustering algorithm on distributed memory multiprocessors", *IBM Research Report*, No. RJ 10134(95009) (Nov. 11, 1998), also available in *Proc. Large-Scale Parallel KDD Systems Workshop* (ACM SIGKDD), Aug. 15–18, 1999.

[Dhillon 1999] Dhillon, I., Modha, D., "Concept decompositions for large sparse text data using clustering", *IBM Research Report*, No. RJ 10147(95022) (July 8, 1999 - declassified March 13, 2000), to appear in *Machine Learning.*

[Echigo et al. 2000] Echigo, T., Kurokawa, M., Tomita, A., Miyamori, H., Iisaku, S., "Video enrichment: retrieval and enhanced visualization based on behaviors of objects", *Proc. Fourth Asian Conference on Computer Vision* (ACCV2000), Taipei, Taiwan (Jan. 8–11, 2000), 364–369.

[Eichmann, Ruiz, Srinivasan 1998] Eichmann, D., Ruiz, M., Srinivasan, P., "Cross-language information retrieval with the UMLS Metathesaurus", *Proc. ACM Special Interest Group on Information Retreival (SIGIR)*, ACM Press, NY (1998), 72–80.

[Ester et al. 1995a] Ester, M., Kriegel, H.-S., Sander, J., Xu, X., "A density-based algorithm for discovering clusters in large spatial databases with noise", *Proc. Second Int'l. Conference* *on Knowledge Discovery in Data Bases and Data Mining*, AAAI Press, Menlo Park, CA (1995).

[Ester et al. 1995b] Ester, M., Kriegel, H.-S., Xu, X., “A database interface for clustering in large spatial databases”, *Proc. First Int'l. Conference on Knowledge Discovery in Data Bases and Data Mining*, AAAI Press, Menlo Park, CA (1995).

[Ester et al. 1995c] Ester, M., Kriegel, H.-S., Xu, X., “Focusing techniques for efficient class identification”, *Proc. Fourth Int'l. Symp. of Large Spatial Databases* (1995).

[Etzioni, Weld 1995] Etzioni, O., Weld, D., “Intelligent agents on the Internet: fact, fiction and forecast”, preprint, Univ. of Washington, Seattle (May 30, 1995):
www.cs.washington.edu/homes/etzioni/

[Faloutsos 1996] Faloutsos, C., *Searching Multimedia Databases by Content*, Kluwer Academic, Boston, MA (1996).

[Faloutsos, Lin 1995] Faloutsos, C., Lin, K.-I., “FastMap: a fast algorithm for indexing, datamining and visualization of traditional and multimedia datasets”, *Proc. of ACM SIGMOD Conference* (1995), 163–174.

[Faloutsos, Oard 1995] Faloutsos, C., Oard, D., “A survey of information retrieval and filtering methods”, *Technical Report*, Dept. Computer Science, Univ. of Maryland, College Park, No. CS-TR-3514 (Aug. 1995).

[Feldman 1998] Feldman, S., “Web search services in 1998: trends and challenges”, *Information Today*, 9, 9 (June 1998): www.infotoday.com/searcher/jun/story2.htm

[Felt, Scales] Felt, E., Scales, J., Analysis of Web robots, home page:
www.wsulibs.wsu.edu/general/robots.htm

[Ferguson 1999] Ferguson, A., “Auction nation”, *Time* (Dec. 27, 1999 – Jan. 3, 2000).

[Finin et al. 1998] Finin, T., Nicholas, C., Mayfield, J., “Software Agents for Information Retrieval”, short course notes, *Third ACM Conference on Digital Libraries*, Pittsburg, PA, ACM Press, NY (June 24–27, 1998).

[Fisher 1995] Fisher, D., “Iterative optimization and simplification of hierarchical clusterings”, *Technical Report*, Dept. Computer Science, Vanderbilt Univ., TN 37235 (1995).

[Flickner et al. 1997] Flickner, M. et al., “Query by image and video content: the QBIC System”, Chapter 1 in Maybury, M. (ed.), *Intelligent Multmedia Information Retrieval*, MIT Press, Cambridge, MA (1997), 7–22.

[Flynn 1996] Flynn, L., “Desperately seeking surfers; Web programmers try to alter search engines' results”, *New York Times* (Nov. 11, 1996), C5.

[Frakes, Baeza-Yates 1992] Frakes, W., Baeza-Yates, R. (eds.), *Information Retreival*, Prentice-Hall, Englewood Cliffs, NJ (1992).

[Frisse 1988] Frisse, M., "Searching for information in a hypertext medical handbook", *Communications of the ACM*, 31, 7 (1988), 880–886.

[Gilbert 1997] Gilbert, D., "Intelligent agents: the right information at the right time", *IBM Report*, IBM Corporation, Research Triangle Park, NC, USA (May 1997).

[Gloor, Dynes 1998] Gloor, P., Dynes, S., "Cybermap: visually navigating the Web", *Journal of Visual Languages and Computing*, 9, 3, (June 1998), 319–336.

[Golub, Van Loan 1996] Golub, G., Van Loan, C., *Matrix Computations*, third ed., John Hopkins Univ. Press, Baltimore, MD (1996).

[Gravano 1997] Gravano, L., "Querying Multiple Document Collections Across the Internet", Ph.D. Thesis, Dept. Computer Science, Stanford Univ. (Aug. 1997).

[Gudivada et al. 1997] Gudivada, V., Raghavan, V., Grosky, W., Kasaanagottu, R., "Information retrieval on the World Wide Web", *IEEE Internet Computing* (Oct.-Nov. 1997), 58–68.

[Guglielmo 1997] Guglielmo, C., *Upside Today* (on-line) (Sept. 4, 1997):
inc.com/cgi-bin/tech_link.cgi?url=http://www.upside.com

[Guha et al. 1998] Guha, S., Rastogi, R., Shim, K., "Cure: an efficient clustering algorithm for large databases", *Proc. ACM SIGMOD Conference on Management of Data* (June 1998).

[Hawking et al. 1999] Hawking, D., Craswell, N., Thistlewaite, P., Harman, D., "Results and challenges in Web search evaluation", *Computer Networks* (special issue: Proc. 8th World Wide Web Conference), 31 (1999), 1321–1330.

[Hearst 1995] Hearst, M., "TileBars: visualization of term distribution information in full text information access", *Proc. ACM SIGCHI, CHI'95*, Denver, Colorado, ACM, NY (May 7–11): www.acm.org/sigchi/chi95/Electronic/documents/papers/mah_bdy.htm

[Hearst 1997] Hearst, M., "Interfaces for searching the Web", *Scientific American* (March 1997), 68–72.

[Hearst 1999] Hearst, M., "User interfaces and visualization", in Baeza-Yates, R., Ribeiro-Neto, B. (eds.), *Modern Information Retreival*, ACM Press, New York (1999), 257–323.

[Hearst, Pederson 1996] Hearst, M., Pederson, J., "Visualizing information retrieval results: a demonstration of the TileBar interface", *Proc. ACM SIGCHI, CHI'96*, Vancouver, B.C., Canada, ACM Press, NY (Apr. 13–18, 1996):
www.acm.org/sigchi/chi96/proceedings/videos/Hearst/mah?txt.htm

[Henzinger et al. 1999] Henzinger, M., Heydon, A., Mitzenmacher, M., Najork, M., "Measuring index quality using random walks on the Web", *Computer Networks* (special issue: Proc. 8th World Wide Web Conference), 31 (1999), 1291–1303.

[Hernández 1996] Hernández, M., "A generalization of band joins and the merge/purge problem", *Ph.D. Thesis*, Dept. Computer Science, Columbia Univ. (1996).

[Hernández, Stolfo 1995] Hernández, M., Stolfo, S., "The merge/purge problem for large databases", *Proc. ACM SIGMOD Int'l. Conference Management of Data* (May 1995), 127–138.

[Howe, Dreilinger 1997] Howe, A., Dreilinger, D., "SavvySearch: a metasearch engine that learns which search engine to query", *AI Magazine*, 18 (2) (1997), 19–25.

[Huberman, Lukose 1997] Huberman, B., Lukose, R., "Social dilemmas and Internet congestion", *Science*, 277 (1997), 535–537.

[Huberman et al. 1998] Huberman, B., Pirolli, P., Pitkow, J., Lukose, R., "Strong regularities in World Wide Web surfing", *Science*, 280 (Apr. 3, 1998), 95–97.

[Hylton 1996] Hylton, J., *Identifying and merging related bibliographic records*, M.S. Thesis, Dept. Computer Science, MIT. Also available as MIT Lab. for Computer Science, *Technical Report*, No. 678 (1996).

[ICASSP] Int'l. Conference on Acoustics, Speech and Signal Processing (ICASSP), *Proceedings of*, IEEE Press, Piscataway, NJ.

[ICIP] Int'l. Conference on Image Processing (ICIP), *Proceedings of*, IEEE Press, Piscataway, NJ.

[IEEE IS 1996] *IEEE Intelligent Systems*, special issue on intelligent agents (Dec. 1996).

[IEEE IS 1998] *IEEE Intelligent Systems*, special issue on knowledge management (May/June 1998).

[IEEE IS 1999] *IEEE Intelligent Systems*, special issue on intelligent information retrieval (July/Aug. 1999).

[IEEE IC 1998] *IEEE Internet Computing*, the editors of, news and trends section, (Mar./Apr. 1998), 8–9.

[IEEE PAMI 1996] *IEEE Trans. on Pattern Analysis and Machine Intelligence*, special issue on digital libraries: representation and retrieval, 18, 8, (Aug. 1996), 783–79.

[IFIP 1989, 1992] IFIP, *Visual Data Base Systems I and II*, North-Holland and Elsevier, Amsterdam (1989, 1992).

[Iwai et al. 2000] Iwai, Y., Maruo, J., Yachida, M., Echigo, T., Iisaku, S., "A framework for visual event extraction from soccer games", *Proc. Fourth Asian Conference on Computer Vision* (ACCV2000), Taipei, Taiwan (Jan. 8–11, 2000), 222–227.

[Jacoby, Slamecka 1962] Jacoby, J., Slamecka, V., "Indexer consistency under minimal conditions", *Technical Report*, No. RADC TR 62-426, Documentation, Inc., Bethesda, MD, AD-288087 (1962).

[Jacobs et al. 1995] Jacobs, C., Finkelstein, A., Salesin, D., "Fast multiresolution image querying", *Proc. of ACM SIGGRAPH* (1995), 277–286.

[Jain, Dubes 1998] Jain, A., Dubes, R., *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ (1988).

[Kageyama, Takashima 1994] Kageyama, T., Takashima, Y., "A melody retrieval method with hummed melody", *IEICE Trans. D-II*, J77, 8 (Aug. 1994), 1543–1551 (in Japanese).

[Kahle 1997] Kahle, B., "Archiving the Internet" (1997):
www.alexa.com/ brewster/essays/sciam_article.html

[Kephart et al. 1998] Kephart, J., Hanson, J., Levine, D., Grosof, B., Sairamesh, J., Segal, R., White, S., "Emergent behavior in information economies", manuscript (1998).

[Kephart, Hanson, Sairamesh 1998] Kephart, J., Hanson, J., Sairamesh, J., "Price-war dynamics in a free-market economy of software agents", *Proc. ALIFE VI*, Los Angeles, June 26-29 (1998).

[Kleinberg 1998] Kleinberg, J., "Authoritative sources in a hyperlinked environment", *Proc. ACM-SIAM Symp. on Discrete Algorithms* (Jan. 1998), also available as *IBM Research Report*, No. RJ 10076 (May 1997).

[Kobayashi et al. 1999] Kobayashi, M., Dupret, G., King, O., Samukawa, H., Takeda, K., "Multi-perspective retrieval, ranking and visualization of Web data, *Proc. Int'l Symposium on Digital Libraries (ISDL) '99*, Tsukuba, Japan (1999), pp. 159–162.

[Korfhage 1997] Korfhage, R., *Information Storage and Retrieval*, John Wiley and Sons, NY (1997).

[Koster 1995] Koster, M., "Robots in the Web: trick or treat ?", *ConneXions*, 9, 4 (Apr. 1995).

[Koster 1996] Koster, M., "Examination of the standard for robots exclusion" (1996):
info.webcrawler.com/mak/projects/robots/eval.html

[Kurokawa et al. 1999] Kurokawa, M., Echigo, T., Tomita, T., Maeda, J., Miyamori, H., Isisaku, S., "Representation and retrieval of video scene by using object actions and their spatio-temporal relationships, *IEEE International Conference on Image Processing* (ICIP) Kobe, Japan, IEEE Press, piscataway, NJ (Oct. 1999) 26A02.1.

[Lagoze 1996] Lagoze, C. "The Warwick Framework: a container architecture for diverse sets of metadata", *D-Lib Magazine* (July/August 1996):
www.dlib.org/dlib/july96/lagoze/07lagoze.html

[Lamping, Rao, Pirolli 1995] Lamping, J., Rao, R., Pirolli, P., "A focus+content technique based on hyperbolic geometry for visualizing large hierarchies"", *Proc. ACM Human Factors in Computing Systems* (May 1995), 401–408.

[Lawrence, Giles 1998a] Lawrence, S., Giles, C., "Context and page analysis for improved Web search", *IEEE Internet Computing*, 2, 4 (1998) 38–46.

[Lawrence, Giles 1998b] Lawrence, S., Giles, C., "Searching the World Wide Web", *Science*, 280 (Apr. 3, 1998), 98–100.

[Lawrence, Giles 1999a] Lawrence, S., Giles, C., "Searching the Web: general and scientific information access", *IEEE Communications*, 37, 1 (1999), 116–122.

[Lawrence, Giles 1999b] Lawrence, S., Giles, C., "Accessibility of information on the Web", *Nature*, 400 (July 8, 1999), 107–109.

[Lawrence, Giles 1999c] Lawrence, S., Giles, C., "Text and Image Metasearch on the Web", *Proc. Int'l. Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA99)*, CSREA Press (1999), 829–835.

[Leighton, Srivastava 1997] Leighton, H., Srivastava, J., "Precision among World Wide Web search engines: AltaVista, Excite, Hotbot, Infoseek, and Lycos" (1997):
www.winona.msus.edu/library/webind2/webind2.htm

[Letsche, Berry 1997] Letsche, T., Berry, M., "Large-scale information retrieval with latent semantic indexing", submitted to *Information Sciences – Applications* (1997).

[Liao et al. 1992] Liao, H., Osada, M., Shneiderman, B., "A formative evaluation of three interfaces for browsing directories using dynamic queries", *Technical Report*, Dept. Computer Science, Univ. Maryland, College Park, No. CS-TR-2841, CAR-TR-605 (1992).

[Liberatore 1998] Liberatore, K., "Getting to the source: is it real or spam, Ma' am ?", *MacWorld* (July 2, 1997), also available at: *Macworld Online* (Aug. 15, 1998):
macworld.zdnet.com/features/pov.4.4.html

[Licklider 1965] Licklider, J., *Libraries of the Future*, MIT Press, Cambridge, MA (1965).

[Lidsky, Kwon 1997] Lidsky, D., Kwon, R., "Searching the Net", *PC Magazine* (Dec. 2, 1997), 227–258.

[Liechti, Sifer, Ichikawa 1998] Liechti, O., Sifer, M., Ichikawa, T., "Structured graph format: XML metadata for describing Web site structure", *Computer, Networks and ISDN Ssystems*, 30, 1-7 (Apr. 1998), 11-21.

[Losee 1998] Losee, R., *Text Retrieval and Filtering: Analytic Models of Performance*, Kluwer, Boston (1998).

[Lynch 1997] Lynch, C., "Searching the Internet", *Scientific American* (March 1997), 52–56.

[Maarek et al. 1997] Maarek, Y., Jacovi, M., Shtalhaim, M., Ur, S., Zernik, D., Shaul, I., "WebCutter: a system for dynamic and tailorable site mapping", *Computer, Networks, and ISDN Systems*, 29, 8-13 (Sept. 1997), 1269–1279.

[Macskassy et al. 1998] Macskassy, S., Banerjee, A., Davison, B., Hirsh, H., "Human performance on clustering Web pages: a preliminary study", *Proc. Fourth Int'l. Conference on Knowledge Discovery and Data Minng*, AAAI Press, Menlo Park CA (1998), 264–268.

[Manber 1999] Manber, U., "Foreword" in Baeza-Yates, R., Ribeiro-Neto, B. (eds.), *Modern Information Retreival*, ACM Press, New York (1999) v-viii.

[Manber, Smith, Gopal 1997] Manber, U., Smith, M., Gopal, B., "Webglimpse: combining borwsing and searching", *Proc. of USENIX Technical Conference*, Anaheim, USA (Jan 1997), 195–206.

[Manjunath, Ma 1996] Manjunath, B., Ma, W., "Texture features for browsing and retrieval of image data", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, special issue on digital libraries: representation and retrieval, 18, 8 (Aug. 1996), 836–842.

[Manning, Schütze 1999] Manning, C., Schütze, H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA (1999).

[Marchionini 1995] Marchionini, G., *Information Seeking in Electronic Environments*, Cambridge Univ. Press, Cambridge, UK (1995).

[Marczewski, Steinhaus 1958] Marczewski, E., Steinhaus, H., "On a certain distance of sets and the corresponding distance of functions", *Colloquium Mathematicum*, 6 (1958), 319–327.

[Maybury 1997] Maybury, M. (ed.), *Intelligent Multmedia Information Retrieval*, MIT Press, Cambridge, MA (1997).

[Maybury, Wahlster 1998] Maybury, M., Wahlster, W. (eds.), *Readings in Intelligent User Interfaces*, Morgan Kaufmann, San Francisco, CA (1998).

[McKnight, Boroumand 2000] McKnight, L., Boroumand, J., "Pricing Internet services: approches and challenges", *IEEE Computer* (Feb. 2000), 128–129.

[Mereu, Kazman 1996] Mereu, S., Kazman, R., "Audio enhanced 3D interfaces for visually impaired users", *Proc. ACM SIGCHI '96*, Vancouver, B.C., ACM Press, NY (Apr. 13–18, 1996): www.acm.org/sigchi/chi96/proceedings/papers/Mereu/rnk-txt.htm

[Mitchell] Mitchell, S., General Internet resource finding tools, home page:
library.ucr.edu/pubs/navigato.html

[Miyahara et al. 2000] Miyahara, T., Watanabe, H., Tazoe, E., Kamiyama, Y., Takeda, K., *Internet Machine Translation*, Mainichi Communications (2000), in Japanese.

[Modha, Spangler 1999] Modha, S., Spangler, W., "Clustering hypertext with applications to Web searching", *IBM Research Report*, No. RJ 10160 (95035) (Oct. 7, 1999), also available in *Proc. of ACM Hypertext Conference*, San Antonio, TX, May 30 – June 3, 2000.

[Monge, Elkan] Monge, A., Elkan, C., "An efficient domain-independent algorithm for detecting approximately duplicate database records", *preprint*, U.C. San Diego:
{amonge, elkan }@cs.ucsd.edu.

[Monier 1998] Monier, L., "AltaVista CTO Responds" (1998):
www4.zdnet.com/anchordesk/talkback/talkback_13066.html

[Morohashi et al. 1995] Morohashi, M., Takeda, K., Nomiyama, H., Maruyama, H., "Information outlining: - filling the gap between visualization and navigation in digital libraries", *Proc. of Int'l. Symp. on Digital Libraries.* Tsukuba, Japan (1995).

[Munzner, Burchard 1995] Munzner, T., Burchard, P., "Visualizing the structure of the World Wide Web in 3D Hyperbolic Space", *Proc. ACM Symposium on VRML Modeling Language* (Dec. 1995), 33-38.

[Nagao, Hasida 1998] Nagao, K., Hasida, K., "Automatic text summarization based on the global document annotation", *Proc. COLING-ACL '98*, 2, Morgan Kaufmann, San Francisco, CA (1998).

[Nagao et al. 1999] Nagao, K., Hosoya, S., Kawakita, Y., Ariga, S., Shirai, Y., Yura, J., "Semantic transcoding: making the World Wide Web more understandable and reusable by external annotations", unpublished manuscript (1999).

[Navarro 1998] Navarro, G., *Approximate Text Searching*, Ph.D. Thesis, Dept. Computer Science, Univ. of Chile (1998).

[Ng, Han 1994] Ng, R., Han, J., "Efficient and effective methods for spatial data mining", *Proc. of Very Large Data Bases* (1994).

[Niblack et al. 1993] Niblack, W., et al., "The QBIC Project: Query by image by content using color, texture, and shape", in Niblack, W., Jain, R. (eds.), *Proc. of Storage and Retrieval for Image and Video Databases*, Vol. 1908 SPIE Press, Bellingham, WA (1993), 173–187.

[Niblack, Jain 1993, 1994, 1995] Niblack, W., Jain, R. (eds.), *Proc. of Storage and Retrieval for Image and Video Databases*, Vols. 1908, 2185, and 2420, SPIE Press, Bellingham, WA (1993, 1994, 1995).

[Nielsen 1993] Nielsen, J., *Usability Engineering*, Academic Press, New York (1993).

[Nielsen 1999] Nielsen, J., "User interface directions for the Web", *Communications of the ACM*, 42 (Jan. 1997), 65–72.

[Nomiyama et al. 1997] Nomiyama, H., Kushida, T., Uramoto, N., Ioka, M., Kusaba, M., Hong, J.-K., Chigono, A., Itoh, T., Tsuji, M., “Information navigation system for multimedia data”, *IBM Tokyo Research Laboratory, Research Report*, No. RT-0227 (1997).

[Nua 1999] Nua’s Surveys: www.nua.ie/surveys
How many online: www.nua.ie/surveys/how_many_online/index.html
Graphs and charts: www.nua.ie/surveys/analysis/graphs_and_charts/index.html

[Oard 1997a] Oard, D., “Speech-based information retrieval for digital libraries”, *Technical Report*, Dept. Computer Science, Univ. Maryland, College Park, No. CS-TR-3778 (Mar. 1997).

[Oard 1997b] Oard, D., “Cross-language text retrieval research in the USA”, *Proc. the Third ERCIM DELOS Wkshp.*, Zurich, Switzerland (March 1997):
www.glue.umd.edu/õard/research.html

[Oard 1997c] Oard, D., “Serving users in many languages”, *D-Lib Magazine* (Dec. 1997):
www.dlib.org/dlib/december97/oard/12oard.html

[Oard, Dorr 1996] Oard, D., Dorr, B., “A survey of multilingual text retrieval”, *Technical Report*, Dept. Computer Science, Univ. Maryland, College Park, No. CS-TR- 3615 (Apr. 1996).

[Omiecinski, Scheuermann 1990] Omiecinski, E., Scheuermann, P., “A parallel algorithm for record clustering”, *ACM Trans. on Database Systems*, 14, 4 (Dec. 1990) 599–624.

[Oogane, Asakawa 1998] Oogane, T., Asakawa, C., “An interactive method for accessing tables in HTML”, *Proc. Third ACM Conference on Assistive Technologies - ASSETS ’98*, Marina Del Rey, CA, ACM Press, NY (Apr. 15-17, 1998), 126–128.

[Parlett 1998] Parlett, B., *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, PA (1998).

[Perkowitz, Etzioni 1999] Perkowitz, M., Etzioni, O., “Adaptive Web Sites: an AI Challenge”, Univ. of Washington, Seattle (1999): info.cs.vt.edu/

[Pike 1999] Pike, J., “Shocked by search engine indexing” (1999):
www4.zdnet.com/anchordesk/talkback/talkback_11638.html

[Pirkola 1998] Pirkola, A., “The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval”, *Proc. ACM Special Interest Group on Information Retreival (SIGIR)*, ACM Press, NY (1998), 55–63.

[Pirolli et al. 1996a] Pirolli, P., Pitkow, J., Rao, R., “Silk from a sow’s ear: extracting usable structures from the Web”, *Proc. of ACM SIGCHI Conference on Human Factors in Computing* (1996).

[Pirolli et al. 1996b] Pirolli, P., Schank, P., Hearst, M., Diehl, C., "Scatter/Gather browsing communicates the topic structure of a very large text collection", *Proc. ACM SIGCHI, CHI'96* Vancouver, B.C., Canada, ACM Press, NY (Apr. 13–18, 1996):
www.acm.org/sigchi/chi96/proceedings/papers/Pirolli/pp_txt.htm

[Plaisant 1994] Plaisant, C., "Dynamic queries on a health statistics atlas", *Technical Report*, Dept. of Computer Science, Univ. of Maryland, College Park (1994).

[Pringle et al. 1998] Pringle, G., Allison, L., Dowe, D., "What is a tall poppy among Web pages ?", *Proc. Seventh World Wide Web Conference* (1998):
www7.conf.au/programme/fullpapers/1872/com1872.htm

[Preschel 1972] Preschel, B., "Indexer consistency in perception of concepts and choice of terminology", Final Report, School of Library Science, Columbia Univ. (1972).

[Press et al. 1982] Press, W., Teukolsky, S., Vetterling, W., Flannery, B., *Numerical Recipes in C*, 2nd ed., Cambridge Univ. Press, NY (1982).

[Raghavan 1997] Raghavan, P., "Information retrieval algorithms: a survey", *ACM-SIAM Proc. Symp. on Discrete Algorithms* (1997).

[Raman 1996] Raman, T., "Emacspeak – a speech interface", *Proc. ACM SIGCHI, CHI'96*, Vancouver, B.C., Canada, ACM Press, NY (Apr. 13–18, 1996):
www.acm.org/sigchi/chi96/proceedings/papers/Raman/paper.html

[Rao et al. 1993] Rao, R., Pederson, J., Hearst, M., Mackinlay, J., Card, S., Masinter, L., Halvorsen, P., Robertson, G., "Rich interaction in the digital library", *Communications of the ACM*, 36, 4 (Apr. 1993), 29–39.

[Rasmussen 1992] Rasmussen, E., "Clustering algorithms", in Frakes, W., Baeza-Yates, R. (eds.), *Information Retreival*, Prentice-Hall, Englewood Cliffs, NJ (1992), 419–442.

[Ratha et al. 1996] Ratha, N., Kaur, K., Chen, S., Jain, A., "A real-time matching system for large fingerprint databases", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, special issue on digital libraries: representation and retrieval, 18, 8, (Aug. 1996), 799–813.

[Rennison 1994] Rennison, E., "Galaxy of news: an approach to visualizing and understanding expansive news landscapes", *Proc. of ACM Symp. on User Interface Software and Technology* ACM New York (1994), 3–12.

[Resnick 1997] Resnick, P., "Filtering information on the Internet", *Scientific American* (March 1997), 62–64.

[Rivlin et al. 1994] Rivlin, E., Botafogo, R., Shneiderman, B., "Navigating hyperspace: designing a structure-based toolbox", *Communications of the ACM*, 37, 2 (1994), 87–96.

[Robertson, MacKinlay, Card 1991] Robertson, G., MacKinlay, J., Card, S., “Cone Trees: animated 3D visualizations of hierarchical information”, *Proc. ACM Human Factors in Computing Systems* (April/May 1991) 189–194.

[Sakairi 1999] Sakairi, T., ”A site map for visualizing both a Web site’s structure and keywords”, *Proc. IEEE System, Man, and Cybernetics Conference* (SMC’99) (1999), 200-205.

[Salton 1969] Salton, G., “A comparison between manual and automatic indexing methods”, *American Documentation*, 20, 1 (1969), 61–71.

[Salton 1970] Salton, G., “Automatic text analysis”, *Science*, 168 (1970), 335–343.

[Salton 1971] Salton, G., *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, NJ (1971).

[Salton 1989] Salton, G., *Automatic Text Processing*, Addison-Wesley, Reading, Mass (1989).

[Salton, Buckley 1988] Salton, G., Buckley, C., “Term-weighting approaches in automatic text retrieval”, *Information Processing and Management*, 24, 5 (1988), 513–523.

[Salton, McGill 1983] Salton, G., McGill, M., *Introduction to Modern Information Retrieval*, McGraw-Hill, NY (1983).

[Samet, Soffer 1996] Samet, H., Soffer, A., “MARCO: MAp Retrieval by COntent”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, special issue on digital libraries: representation and retrieval, 18, 8 (Aug. 1996), 783–798.

[Schatz 1997] Schatz, B., “Information retrieval in digital libraries: bringing search to the Net”, *Science*, 275 (Jan. 17, 1997), 327–334.

[Schäuble 1999] Schäuble, P., *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*, Kluwer, Boston (1997).

[Schneier, Abdel-Mottaleb 1996] Schneier, M., Abdel-Mottaleb, M., “Exploiting the JPEG compression scheme for image retrieval”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, special issue on digital libraries: representation and retrieval, 18, 8 (Aug. 1996), 849–853.

[Scientific American 1997] *Scientific American*, “The Internet: fulfillling the promise”, special report (March 1997).

[Selberg, Etzioni 1995] Selberg, E., Etzioni, O., “Multiple service search and comparison using the MetaCrawler”, *Proc. of World Wide Web Conference* (1995).

[Selberg, Etzioni 1997] Selberg, E., Etzioni, O., “The MetaCrawler architecture for resource aggregation on the Web”, *IEEE Expert* (1997).

[Shakes et al. 1997] Shakes, J., Langheinrich, M., Etzioni, O., "Dynamic reference sifting: a case study in the homepage domain", *Proc. Sixth Int'l. World Wide Web Conference* (1997), 189–200.

[Shivakumar, García-Molina 1998] Shivakumar, N., García-Molina, H., "Finding near-replicas of documents on the Web", *Proc. Wkshp. on Web Databases*, Valencia, Spain (March 1998).

[Shneiderman 1994] Shneiderman, B., "Dynamic queries for visual information seeking", *Technical Report*, Dept. Computer Science, Univ. Maryland, College Park, No. CAR-TR-655, CS-TR-3022, SRC-TR-93-3 (Sept. 1993, revised Jan. 1994).

[Shneiderman 1994] Shneiderman, B., *Designing the User Interface*, third ed., Addison Wesley Longman, Inc., Reading, MA (1998).

[Silberschatz et al. 1995] Silberschatz, A., Stonebraker, M., Ullman, J., "Database research: achievements and opportunities into the 21st century", *Technical Report*, NSF Wkshp. on the Future of Database Research (May 1995).

[Small 1973] Small, H., "Co-citation in the scientific literature: a new measure of the relationship between two documents", *Journal of the American Society for Information Science*, 24 (1973), 265–269.

[Small, Sweeney 1985] Small, H., Sweeney, E., "Clustering the science citation index using cocitations, Part I: A comparison of methods", *Scientometrics*, 7 (1985), 391–409.

[Smith 1997] Smith, Z., "The truth about the Web: crawling towards eternity", *Web Techniques Magazine* (May 1997):
www.webtechniques.com/features/1997/05/burner/burner.html

[Smith, Chang 1996a] Smith, J., Chang, S.-F., "Searching for images and videos on the World-Wide Web", *IEEE Multimedia Magazine*, Summer 1997, also Columbia Univ., *CU/CTR Technical Report*, No. 459-96-25 (1996):
ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/smith96e.ps
demo: www.ctr.columbia.edu/webseek

[Smith, Chang 1996b] Smith, J., Chang, S.-F., "VisualSEEk: a fully automated content-based image query system", *Proc. ACM Multimedia Conference*, Boston, MA (Nov. 1996):
ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/smith96f.ps
demo: www.ctr.columbia.edu/VisualSEEk

[Smith, Chang 1997] Smith, J., Chang, S.-F., "Querying by color regions using the VisualSEEk content-based visual query system", in Maybury, M. (ed.), *Intelligent Multmedia Information Retrieval*, MIT Press, Cambridge, MA (1997), 23–41 (Chapter 2).

[Smoliar 1994] Smoliar, S., Zhang, H., "Content-based video indexing retrieval", *IEEE Multimedia* (Summer 1994), 62–72.

[Sneath, Sokal 1973] Sneath, P., Sokal, R., *Numerical Taxonomy*, Freeman, San Francisco, CA (1973).

[Soergel 1985] Soergel, D., *Organizing Information*, Academic Press, London, UK, 1985.

[Soffer, Samet 1999] Soffer, A., Samet, H., "Pictorial query specification for browsing through spatially-referenced Image databases", *Journal of Visual Languages and Computing*, to appear: www.cs.umd.edu/ hjs/

[Sparck Jones, Willett 1997] Sparck Jones, K., Willett, P. (eds.), *Readings in Information Retrieval*, Morgan Kaufmann Publishers, Inc., San Francisco, CA (1997).

[SIGCHI] Special Interest Group on Computer-Human Interaction (SIGCHI), the Association for Computing Machinery, home page: www.acm.org/sigchi/

[SIGIR] Special Interest Group on Information Retrieval (SIGIR), the Association for Computing Machinery, home page: www.acm.org/sigir/

[Strategy Alley 1998] Strategy Alley, "White paper on the viability of the Internet for business", (Apr. 29, 1998): www.strategyalley.com/articles/inet1.htm

[Strzalkowski 1999] Strzalkowski, T., *Natural Language Information Retreival*, Kluwer, Boston (1999).

[Syone et al. 1994] Syone, C., Fishkin, K., Bier, E., "The movable filter as a user interface tool", *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, Boston, MA (Apr. 24-28, 1994), 306–312.

[Swets, Weng 1996] Swets, D., Weng, J., "Using discriminant eigenfeatures for image retrieval", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, special issue on digital libraries: representation and retrieval, 18, 8 (Aug. 1996), 831–836.

[Sze, Liao, Lu 2000] Sze, C.-J., Liao, H.-Y., Lu, C.-S., "Shape-based retrieval of fish databse of Taiwan", *Proc. Fourth Asian Conference on Computer Vision* (ACCV2000), Taipei, Taiwan (Jan. 8–11, 2000), 370–375.

[Takeda, Nomiyama 1997] Takeda, K., Nomiyama, H., "Information outlining and site outlining", *Proc. of Int'l. Symp. on Digital Libraries*, Tsukuba, Japan (1997).

[Tetranet 1998] Tetranet Software, Inc., *Wisebot* (1998):
www.tetranetsoftware.com/products/wisebot.htm

[Tufte 1983] Tufte, E., *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut (1983).

[van Rijsbergen 1977] van Rijsbergen, C., "A theoretical basis for the use of cooccurrence data in information retrieval", *Journal of Documentation*, 33, 2 (June 1977), 106–119.

[van Rijsbergen 1979] van Rijsbergen, C., *Information Retrieval*, second ed., Butterworths, Boston (1979).

[Voorhees 1986a] Voorhees, E., *The Effectiveness and Efficiency of Implementing Agglomerative Hierarchic Clustering Algorithms for use in Document Retrieval*, Ph.D. Thesis, Cornell Univ. (1986).

[Walker et al. 1998] Walker, J., Case, T., Jorasch, J., Sparico, T., "Method, apparatus, and program for pricing, selling, and exercising options to purchase airline tickets", U.S. Patent no. 5797127, filed Dec. 31, 1996, issued Aug. 18, 1998.

[Walker, Sparico, Case 1999] Walker, J., Sparico, T., Case, T., "Method and apparatus for the sale of airline-specified flight tickets", U.S. Patent no. 5897620, filed July 8, 1997, issued April 27, 1999.

[Watanabe, Takeda 1998] Watanabe, H., Takeda, K., "A pattern-based machine translation system extended by example-based processing", *Proc. COLING-ACL '98*, 2, Morgan Kaufmann, San Francisco, CA (1998), 1369–1373.

[Webster, Paul 1996] Webster, K., Paul, K., "Beyond surfing: tools and techniques for searching the Web" (Jan. 1996): magi.com/m̃melick/it96jan.htm

[Westera 1996] Westera, G., "Robot-driven search engine evaluation overview" (October 1996): www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/

[White, McCain 1989] White, H., McCain, K., "Bibliometrics", *Annual Review Information Science and Technology*, Elsevier, Amesterdam, Holland (1989), 119–186.

[Willett 1988] Willett, P., "Recent trends in hierarchical clustering: a critical review", *Information Processing and Management*, 24 (1988), 577-597.

[Williamson, Shneiderman 1992] Williamson, C., Shneiderman, B., "The Dynamic HomeFinder: evaluating dynamic queries in a real-estate information exploration system", *Proc. ACM Special Interest Group on Information Retreival (SIGIR)*, ACM Press, NY (1992), 339–346.

[Williams 1984] Williams, M., "What makes RABBIT run ?", *Int'l. Journal of Man-Machine Studies*, 21 (1984), 333–352.

[Wise et al. 1995] Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., "Visualizing the non-visual: spatial analysis and interaction with information from text documents", *Proc. Information Visualization Symp. '95*, IEEE Computer Society Press, Piscataway, NJ (1995), 51–58.

[Witten, Moffat, Bell 1994] Witten, I., Moffat, A., Bell, T., *Managing Gigabytes*, Van Nostrand Reinhold, NY (1994).

[W3] Int'l World Wide Web Conferences (IW3C):
www.w3.org/Conferences/Overview-WWW.html

[Wu, Narasimhalu 1994] Wu, J., Narasimhalu, A., "Identifying faces using multiple retrievals", *IEEE Multimedia* (Summer 1994), 27–38.

[Zamir, Etzioni 1998] Zamir, O., Etzioni, O., "Web document clustering: a feasibility demonstration", *Proc. ACM Special Interest Group on Information Retreival (SIGIR)*, ACM Press, NY (1998), 46–54.

[Zamir et al. 1997] Zamir, O., Etzioni, O., Madani, O., Karp, R., "Fast and intuitive clustering of Web documents", *Proc. Third Int'l Conference on Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, CA (1997), 287–290.

[Zhang et al. 1996] Zhang, T., Ramakrishnan, R., Livny, M., "Birch: an efficient data clustering method for large databases", *Proc. SIGMOD 96* (1996).

[Zhong, Chang 1997] Zhong, D., Chang, S.-F., "Video object model and segmentation for content-based video indexing," *Proc. IEEE Int'l Conference on Circuits & Systems*, Hong Kong, special session on Networked Multimedia Technology & Application (June 1997): ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/97/zhong97a.ps