Appears in *IEEE Communications*, 37(1):116–122, 1999.

Searching the Web: General and Scientific Information Access

Steve Lawrence and C. Lee Giles
NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
Phone: 609 951 2676 (lawrence) 2642 (giles) Fax: 609 951 2482
{lawrence,giles}@research.nj.nec.com

Abstract

The World Wide Web has revolutionized the way that people access information, and has opened up new possibilities in areas such as digital libraries, general and scientific information dissemination and retrieval, education, commerce, entertainment, government, and health care. There are many avenues for improvement of the Web, for example in the areas of locating and organizing information. Current techniques for access to both general and scientific information on the Web provide much room for improvement – search engines do not provide comprehensive indices of the Web and have difficulty in accurately ranking the relevance of results. Scientific information on the Web is very disorganized. We discuss the effectiveness of Web search engines, including results that show that the major Web search engines cover only a fraction of the "publicly indexable Web". Current research into improved searching of the Web is discussed, including new techniques for ranking the relevance of results, and new techniques in metasearch that can improve the efficiency and effectiveness of Web search. The creation of digital libraries incorporating autonomous citation indexing is discussed for improved access to scientific information on the Web.

1 Introduction

The World Wide Web is revolutionizing the way that people access information, and has opened up new possibilities in areas such as digital libraries, general and scientific information dissemination and retrieval, education, commerce, entertainment, government, and health care. The amount of publicly available information on the Web is increasing rapidly (Lawrence and Giles, 1998*b*). The Web is a gigantic digital library, a searchable 15 billion word encyclopedia (Barrie and Presti, 1996). It has stimulated research and development in information retrieval and dissemination, and fostered search engines such as AltaVista. These new developments are not limited to the Web, and can enhance access to virtually all forms of digital libraries.

The revolution that the Web has brought to information access is not so much due to the availability of information (huge amounts of information has long been available in libraries and elsewhere), but rather the increased efficiency of accessing information, which can make previously impractical tasks practical. There

are many avenues for improvement in the efficiency of accessing information on the Web, for example in the areas of locating and organizing information.

This article discusses general and scientific information access on the Web, and many of our comments are applicable to digital libraries in general. The effectiveness of Web search engines is discussed, including results that show that the major search engines cover only a fraction of the "publicly indexable Web" (the part of the Web which is considered for indexing by the major engines, which excludes pages hidden behind search forms, pages with authorization requirements, etc.). Current research into improved searching of the Web is discussed, including new techniques for ranking the relevance of results, and new techniques in metasearch that can improve the efficiency and effectiveness of Web search.

The amount of scientific information and the number of electronic journals on the Internet continues to increase. Researchers are increasingly making their work available online. This article also discusses the creation of digital libraries of the scientific literature incorporating autonomous citation indexing. The autonomous creation of citation indices is possible today, and can improve access to scientific information on the Web or in other digital libraries of scientific articles.

2 Web Search

One of the key aspects of the World Wide Web that makes it a valuable information resource is that the full text of documents can be searched using Web search engines such as AltaVista and HotBot. Just how effective are the Web search engines? The following sections discuss the effectiveness of current engines and current research into improved techniques.

2.1 Comprehensiveness and Recency of the Web Search Engines

This section considers the effectiveness of the major Web search engines in terms of comprehensiveness and recency. We provide results on the size of the Web, the coverage of each search engine, and the freshness of the search engine databases. These results show that none of the search engines covers more than about one third of the publicly indexable Web, and that the freshness of the various databases varies significantly.

Typical quotes regarding the coverage and recency of the major search engine databases include: *"If you can't find it using AltaVista search, it's probably not out there"* (Seltzer, Ray and Ray, 1997), *"[With AltaVista] you can find new information just about as quickly as it's available on the Web"* (Seltzer et al., 1997), and *"HotBot is the first search robot capable of indexing and searching the entire Web"* (Inktomi, 1997). However, the World Wide Web is a distributed, dynamic, and rapidly growing (Lawrence and Giles, 1998*b*) information resource that presents difficulties to traditional information retrieval technologies. Traditional information retrieval software was designed for different environments and has typically been used for indexing a static collection of directly accessible documents. The nature of the Web brings up questions such

as: can the centralized architecture of the search engines keep up with the increasing number of documents on the Web? Can they update their databases regularly to detect modified, deleted, and relocated information? Answers to these questions impact on the best methodology to use when searching the Web, and on the future of Web search technology.

We performed a study of the comprehensiveness and recency of the major Web search engines in December 1997 by analyzing the responses of AltaVista, Excite, HotBot, Infoseek, Lycos, and Northern Light for 575 queries that were made by employees at the NEC Research Institute (Lawrence and Giles, 1998*b*). Search engines rank documents differently and can return documents that do not contain the query terms (e.g. pages with morphological variants or synonyms). Therefore, we only considered queries for which we could download the full-text of every document that each engine reports as matching the query. Documents were only counted if they could be downloaded and they contained the query terms. We handled other important details such as the normalization of URLs, and capitalization and morphology details (full details can be found in (Lawrence and Giles, 1998*b*)).

Table 1 shows the estimated coverage of the search engines, which varies by an order of magnitude. This variation is much greater than would be expected from considering the number of pages that each engine reports to have indexed. The variation may be explained by differences in indexing or retrieval technology between the engines (e.g. an engine would appear to be smaller if it only indexed part of the text on some pages), or by differences in the kinds of pages indexed (our study used mostly scientific queries which may not be covered as well if an engine focuses more on well-connected, popular pages).

We estimated a lower bound on the size of the publicly indexable Web to be 320 million pages. In order to produce this estimate, we analyzed the overlap between pairs of engines (Lawrence and Giles, 1998*b*). Consider two engines a and b. Using the assumption that each engine samples the Web independently, the quantity $\frac{n_o}{n_b}$, where n_o is the number of documents returned by both engines and n_b is the number of documents returned by engine b, is an estimate of the fraction of the indexable Web, p_a, covered by engine a. The size of the indexable Web can then be estimated with s_a/p_a where s_a is the number of pages indexed by engine a. This technique is limited because the engines do not choose pages to sample independently, e.g. they all allow pages to be registered, and they are typically biased towards indexing more popular or well-connected pages. To estimate the size of the Web we used the overlap between the largest two engines where the independence assumption is more valid (the larger engines can index more of the non-registered and less popular pages). Some dependence between the sampling of the engines remains between the largest two engines, and therefore this estimate is a lower bound. Using this estimate of the size of the Web, we found that no engine indexes more than about one third of the indexable Web. We also found that combining the results of the six engines returned approximately 3.5 times more documents on average when compared to using only one engine.

Recall that the queries used in the study were from the employees of the NEC Research Institute. Most of the employees are scientists, and scientists tend to search for less “popular”, or harder to find information. This is beneficial when estimating the size of the Web as above. However, the search engines are typically biased

towards indexing more "popular" information. Therefore the coverage of the search engines is typically better for more popular information.

There are a number of possible reasons why the major search engines do not provide comprehensive indices of the Web: the engines may be limited by network bandwidth, disk storage, computational power, scalability of their indexing and retrieval technology, or a combination of these items (despite claims to the contrary (Steinberg, 1996)). Because Web pages are continually added and modified, a truly comprehensive index would have to index all pages simultaneously, which is not currently possible. Furthermore, there may be many pages that contain no links to them, making it difficult for the search engines to know that the pages exist.

We also looked at the percentage of dead links returned by the search engines, which is related to how often the engines update their databases. Intuitively, it is possible for a tradeoff to exist between the comprehensiveness and the freshness of a search engine – it should be possible to check for modified documents and update the index more rapidly if the index is smaller. Some evidence of such a tradeoff was found – the most comprehensive engine had the largest percentage of dead links, and the least comprehensive engine had the smallest percentage of dead links (Table 1 shows the percentage of invalid links for each search engine). However, we found that the rating of the engines in terms of the percentage of dead links varies greatly over time. This provides evidence that the search engines may not be very regular in their indexing processes, e.g., an engine might suspend the processing of new pages for a period of time during upgrades.

Search Engine	HotBot	AltaVista	Northern Light	Excite	Infoseek	Lycos
Coverage WRT Estimated Web Size	34%	28%	20%	14%	10%	3%
Percentage of Dead Links Returned	5.3%	2.5%	5.0%	2.0%	2.6%	1.6%

Table 1. Estimated coverage of each engine with respect to the estimated size of the Web, and the percentage of invalid links returned by each engine (from 575 queries performed during 12/15/97 – 12/17/97). Note that these results are specific to the particular queries performed (typical queries made by scientists), and the state of the engine databases at the time they were performed. The coverage results may be partly due to different indexing and retrieval techniques rather than different database sizes.

How can this knowledge of the effectiveness of the search engines be used to improve Web search? The coverage investigations indicate that the coverage of the Web engines is much lower than commonly believed, and that the engines tend to index different sets of pages. This indicates that when searching for less popular information, it can be very useful to combine the results of multiple engines. The freshness investigations indicate that it is difficult to predict ahead of time which search engine will be the best engine to use when looking for recent information. Therefore it can also be very useful to combine the results of multiple engines when looking for recent information. There are other ways to compare the search engines besides comprehensiveness and recency. For example, how well the engines rank the relevance of results (discussed in the next section), and features of the query interface.

2.2 Research in Web Search

Research into technology for searching the Web is abundant, which is not surprising considering that the existence of full-text search engines is one of the major differences between the Web and previous means of accessing information. The following sections look specifically at some of the recent research: improved methods for ranking pages that utilize the graph structure of the Web, a metasearch technique which can improve the efficiency of Web search by downloading matching pages in order to extract query term context and analyze the pages, and "softbots" which can be used to locate pages that may not be indexed by any of the engines.

2.2.1 Page Relevance

A common complaint against search engines is that they return too many pages, and that many of them have low relevance to the query. This has been used as an argument for not providing comprehensive indices of the Web ("people are already overloaded with too much information"). However, a search engine could be more comprehensive while still returning the same set of pages first. One of the main problems is that the search engines do not rank the relevance of results very well. Research search engines such as Google (Brin and Page, 1998) and LASER (Boyan, Freitag and Joachims, 1996) promise improved ranking of results. These engines make greater use of HTML structure and the graph formed by hyperlinks in order to determine page relevancy compared to the major Web search engines. For example, Google uses a ranking algorithm called PageRank that iteratively uses information from the number of pages pointing to each page (which is related to the popularity of the pages). Google also uses the text in links to a page as descriptors of the page (the links often contain better descriptions of the page than the pages themselves). Another engine with a novel ranking measure is Direct Hit (`http://www.directhit.com`), which is typically good for common queries. Direct Hit ranks results for a given query according to the number of times previous users have clicked on the pages, i.e. the more popular pages are ranked higher.

Kleinberg (1998) has presented a method for locating two types of useful pages – *authorities*, which contain a lot of information about a topic, and *hubs*, which contain a large number of links to pages about the topic. The underlying principle is the following: good hub pages point to many good authority pages, and a good authority page is pointed to by many good hub pages. A locally iterative process can be used to find hubs and authorities (Kleinberg, 1998). Future search engines may use this method to classify hub and authority pages and to rank the pages within these classes.

2.2.2 Metasearch

Limitations of the search services have led to the introduction of metasearch engines (Selberg and Etzioni, 1995). A metasearch engine searches the Web by making requests to multiple search engines such as

AltaVista or HotBot. The primary advantages of current metasearch engines are the ability to combine the results of multiple search engines and the ability to provide a consistent user interface for searching these engines.

The idea of querying and collating results from multiple databases is not new. Companies like PLS, Lexis-Nexis, DIALOG, and Verity have long since created systems that integrate the results of multiple heterogeneous databases (Selberg and Etzioni, 1995). Many other Web metasearch services exist such as the popular and useful MetaCrawler service (Selberg and Etzioni, 1995). Services similar to MetaCrawler include SavvySearch and Infoseek Express.

Metasearch engines can introduce their own deficiencies, e.g. they can have difficulty ranking the list of results. If one engine returns many low relevance documents, these documents may make it more difficult to find relevant pages in the list. Most of the metasearch engines on the Web also limit the number of results that can be obtained, and typically do not support all of the features of the query languages for each engine.

The NEC Research Institute has been developing an experimental metasearch engine called *Inquirus* (Lawrence and Giles, 1998*a*). Inquirus was motivated by problems with current metasearch engines, as well as the poor precision, limited coverage, limited availability, limited user interfaces, and out of date databases of the major Web search engines. Rather than work with the list of documents and summaries returned by search engines, as current metasearch engines typically do, Inquirus works by downloading and analyzing the individual documents. Inquirus makes improvements over existing engines in a number of areas, e.g. more useful document summaries incorporating query term context, identification of both pages that no longer exist and pages that no longer contain the query terms, improved detection of duplicate pages, progressive display of results, improved document ranking using proximity information (because Inquirus has the full-text of all pages it avoids the ranking problem with standard metasearch engines), dramatically improved precision for certain queries by using specific expressive forms, and quick jump links and highlighting when viewing the full documents.

One of the fundamental features of Inquirus is that it analyzes each document and displays local context around the query terms. The benefit of displaying the local context, rather than an abstract or query-insensitive summary of the document, is that the user may be able to more readily determine if the document answers his or her specific query (without repeatedly clicking and waiting for pages to download). A user can therefore find documents of high relevance by quickly scanning the local context of the query terms. This technique is simple, but it can be very effective, especially for Web search where the database is very large, diverse, and poorly organized.

A study by Tombros (1997) shows that query-sensitive summaries can improve the efficiency of search. Tombros considered the use of query-sensitive summaries and performed a user study that showed that users working with the query-sensitive summaries had a higher success rate. The query-sensitive summaries allowed users to perform relevance judgments more accurately and more rapidly, and greatly reduced the need to refer to the full text of documents.

One interesting feature of Inquirus is the *Specific Expressive Forms* (SEF) search technique. The Web is highly redundant and techniques that trade recall (the fraction of all relevant documents that are returned) for improved precision (the fraction of returned documents that are relevant) are often useful. The SEF search technique transforms queries in the form of a question into specific forms for expressing the answer. For example, the query `What does NASDAQ stand for?` is transformed into the query `‘‘NASDAQ stands for’’ ‘‘NASDAQ is an abbreviation’’ ‘‘NASDAQ means’’`. Clearly the information may be contained in a different form to these three possibilities, however if the information does exist in one of these forms, there is a higher likelihood that finding these phrases will provide the answer to the query. For many queries, the answer might exist on the Web but not in any of the specific forms used. However our experiments indicate that the method works well enough to be effective for certain queries.

Inquirus is surprisingly efficient. Inquirus downloads search engine responses and Web pages in parallel and typically returns the first result faster than the average response time of a search engine.

In summary, metasearch techniques can improve the efficiency of Web search by combining the results of multiple search engines, and by implementing functionality which is not provided by the underlying engines (such as extracting query term context and filtering dead links). The Inquirus metasearch prototype at the NEC Research Institute has shown that downloading and analyzing pages in real-time is feasible. Inquirus, like other meta engines and various Web tools, relies on the underlying search engines, which provide important and valuable services. Wide scale use of this or any metasearch engine would require an amicable arrangement with the underlying search engines. Such arrangements may include passing through ads or micro-payment systems.

2.3 Improving Web Search

Users tend to make queries that result in poor precision. About 70% of queries to Infoseek contain only one term (Harry Motro, Infoseek CEO, CNBC, May 7, 1998). About 40% of queries made by the employees of the NEC Research Institute to the Inquirus engine contain only one term. In information retrieval, there is typically a tradeoff between precision and recall. Simple queries, such as single term queries, can return thousands or millions of documents. Unfortunately, ranking the relevance of these documents is a difficult problem, and the desired documents may not appear near the top of the list. One way to improve the precision of results is to use more query terms, and to tell the search engines that relevant documents must contain certain terms (required terms). Other ways include using phrases or proximity (e.g. searching for specific phrases rather than single terms), using constraints offered by some search engines such as date ranges and geographic restrictions, or using the refinement features offered by some engines (e.g. AltaVista offers a refine function and Infoseek allows subsequent searches within the results set of previous searches).

Another alternative is to combine available search engines with automated online searching. One example is the Internet “softbot” (Etzioni and Weld, 1994). The softbot transforms queries into goals and uses a planning algorithm (that has extensive knowledge of the information sources) in order to generate a sequence

of actions that satisfies the goal. AHOY! is a successful softbot that locates homepages for individuals (Etzioni and Weld, 1994). Shakes et al. performed a study where they searched for the homepages of 582 researchers, and AHOY! was able to locate more homepages than MetaCrawler (which located more homepages than HotBot or AltaVista). AHOY! also provided greatly improved precision.

More comprehensive and more relevant results may also be possible using a search engine that specializes in a particular area, for example Excite NewsTracker specializes in indexing news sites, and OpenText Pinstripe specializes in indexing business sites. Because there are fewer pages to index, the engines may be able to be more comprehensive within their area, and may also be able to update the index more regularly. When searching for popular information, directories constructed by hand, such as Yahoo's directory, can be very useful because fewer low relevance results are returned.

In summary, there exist several ways of improving on the major Web search engines, depending on the type of information desired. For harder to find information, metasearch and softbots can improve coverage. If the topic being queried is covered by one of the more specialized engines, these engines can be used, and they often provide more comprehensive and up-to-date indices within their specialty compared to the general Web search engines.

3 Scientific Information Retrieval

Immediate access to scientific literature has long been desired by scientists, and the Web search engines have made a large and growing body of scientific literature and other information resources accessible within seconds. Advances in computing and communications, and the rapid rise of the the Web, have led to the increasingly widespread availability of online research articles, as well as a simple to use Web version of the Institute for Scientific Information's ® (ISI) *Science Citation Index* ® – the *Web of Science* ®. The Web is changing the way that researchers locate and access scientific publications. Many print journals now provide access to the full-text of articles on the Web, and the number of online journals was about 1,000 in 1996 (Taubes, 1996). Researchers are increasingly making their work available on their homepages or in technical report archives.

3.1 Availability

A lot of scientific literature is copyrighted by the authors or publishers and is not generally available on the "publicly indexable Web." However, the amount of scientific material available on the publicly indexable Web is growing. Some journals owned by societies such as IEEE (the largest technical/scientific society) and ACM are permitting their papers to be placed on the author's Web sites as long as the proper copyright notices are posted. Some private publishers, MIT Press for example, are doing the same. Some publishers

permit prepublication Web access but do not allow the posting of the final version of papers. We predict that more and more papers will be available on the publicly indexable Web in the future.

We used six major Web search engines to search for the papers in a recent issue of *Neural Computation*, after the table of contents was released, but before we obtained our copy of the journal. We found that about 50% of the papers were available on the homepages of the authors. As mentioned before, the coverage of any one search engine is limited. The simplest means of improving the chances of finding a particular scientist or paper on the publicly indexable Web is to combine the results of multiple engines as is done with metasearch engines such as MetaCrawler.

Although more and more scientific papers are being made available on the publicly indexable Web, these papers are spread throughout researcher and institution homepages, technical report archives, and journal sites. The Web search engines do not make it easy to locate these papers because the search engines typically do not index Postscript or PDF documents, which account for a large percentage of the available articles. The next section introduces a technique for organizing and indexing this literature.

3.2 Digital Libraries and Citation Indexing

The Web offers the possibility of providing easy and efficient services for organizing and accessing scientific information. A citation index is one such service. Citation indices (Garfield, 1979) index the citations that an article makes, linking the articles with the cited works. Citation indices were originally designed for literature search, allowing a researcher to find subsequent articles that cite a given article. Citation indices are also valuable for other purposes, including evaluation of articles, authors, etc., and the analysis of research trends. The most popular citation indices of academic research are produced by The Institute for Scientific Information. One such index, the *Science Citation Index*, is intended to be a practical, cost-effective tool for indexing the significant scientific journals. Unfortunately, the ISI databases are expensive and not available to all researchers. Much of the expense is due to the manual effort required during indexing.

The rise of the Internet and the Web has lead to proposals for online digital libraries that incorporate citation indexing. For example, Cameron proposed a "universal, [Internet-based,] bibliographic and citation database linking every scholarly work ever written" (Cameron, 1997). Such a database would be highly "comprehensive and up-to-date", making it a powerful tool for academic literature research, and for the production of statistics as with traditional citation indices. However, Cameron's proposal presents significant difficulty for implementation, and requires authors or institutions to provide citation information in a specific format.

The NEC Research Institute is working on a digital library of scientific publications that creates a citation index autonomously (using *Autonomous Citation Indexing*), without the requirement of any additional effort on the part of the authors or institutions, and without any manual assistance (Giles, Bollacker and Lawrence, 1998). An Autonomous Citation Indexing (ACI) system autonomously extracts citations, iden-

tifies identical citations that occur in different formats, and identifies the context of citations in the body of articles. As with traditional citation indices like the *Science Citation Index*, ACI allows literature search using citation links, and the ranking of papers, journals, authors, etc. by the number of citations. In comparison with traditional citation indexing systems, ACI has both disadvantages and advantages. The disadvantages include lower accuracy (which is expected to be less of a disadvantage over time). However, the advantages are significant and include no manual effort required for indexing, resulting in a corresponding reduction in cost and increase in availability, and literature search based on the context of citations – given a particular paper of interest, an ACI system can display the context of how the paper is cited in subsequent publications. The context of citations can be very useful for efficient literature search and evaluation. ACI has the potential for broader coverage of the literature because human indexers are not required, and can provide more timely feedback and evaluation by indexing items such as conference proceedings and technical reports. Overall, ACI can improve scientific communication, and facilitates an increased rate of scientific dissemination and feedback.

ACI is ideal for operation on the Web – new articles can be automatically located and indexed when they are posted on the Web or announced on mailing lists, and an efficient interface for browsing the articles, citations, and the context of the citations can be created. Part of the benefit of autonomous citation indexing is due to the ability to format and organize information on demand using a Web interface to the citation index. Figure 1 shows an example of the output from the NEC Research Institute's prototype autonomous citation indexing digital library system: CiteSeer. Figure 2 shows an example of how an ACI system can extract the context of citations to a given paper and display them for easy browsing. Note that finding and extracting the context of citations to a given paper could previously be done by using traditional citation indices and manually locating and searching the citing papers – the difference is that the automation and Web interface make the task far more efficient, and thus practical where it may not have been before.

Digital libraries incorporating citation indexing can be used to organize the scientific literature, and help with literature search and evaluation. A "universal citation database" which accurately indexes all literature would be ideal, but is currently impractical because of the limited availability of articles in electronic form, and the lack of standardization in citation practices. However, CiteSeer shows that it is possible to organize and index the subset of literature available on the Web, and to autonomously process free-form citations with reasonable accuracy. As long as there is a significant portion of publishing through the Web, be it the publicly indexable Web, or the subscription-only Web, there is great value in being able to prepare citation indices from the machine readable material. Citation indices may appear that index from both parts of the Web. Access to the full-text of articles may be done openly or by subscription, depending on how the Web and the publication business evolves. Citation indices for subscription-only data may be offered by the publisher or performed by a third-party that has an agreement with the publisher.

Searching for "**simulated annealing**" in **Machine Learning [small test index]** (13828 documents 278202 citations total).

1218 citations found

Click on the [Context] links to see the citing documents and the context of the citations.

Citations (self)	Article
196	S.Kirkpatrick, C.D.Gelatt, and M.P.Vecchi. " *Optimization by* ***simulated annealing***", Science, vol. 220, pp. 671-680, 1983. [Context] [Check]
49 (1)	D.S. Johnson, C.R. Aragon, L.A. McGeoch, and C. Schevon. *Optimization by* ***simulated annealing****: An experimental evaluation*. Technical report. Bell Labs preprint. [Context] [Check]
39	E. Aarts and J. Korst. ***Simulated Annealing*** *and Boltzmann Machines*. John Wiley and Sons, 1989. [Context] [Check]

[... section deleted ...]

Figure 1. An autonomous citation indexing system can group variant forms of citations to the same paper (citations can be written in many different formats), and rank search results by the number of citations. This example shows the results of a search for citations containing the phrase “simulated annealing” in a small test database of the machine learning literature (only a subset of the machine learning literature on the Web). Searching for citations to papers by a given author can also be performed (including secondary authors). The `[Context]` links show the context of the individual citations, a sample of which can be seen in figure 2. The `[Check]` links show the individual citations in each group and can be used to check for errors in the citation grouping.

4 The Future of Web Search and Digital Libraries

What is the future of Web, Web search and digital libraries? Improvements in technology will enable new applications. Computational and storage resources will continue to improve. Bandwidth is likely to increase significantly as technology advances and the following positive spiral works: more people are becoming connected to the Internet as it becomes easier to use, more popular, and as new access mechanisms are introduced (e.g. cable modems and xDSL). This provides incentive for the infrastructure companies to make more investment in the backbone, improving bandwidth. More investment in the backbone improves access so that more people want to be connected.

Will the fraction of the Web covered by the major search engines increase? Some search engines are focusing on indexing Web pages that satisfy the majority of searches, as opposed to trying to catalog all of the Web. However, there are still some engines that aim to index the Web comprehensively. Improvements in indexing technology and computational resources will allow the creation of larger indices. Nevertheless, it is unlikely to become economically practical for a single search engine to index close to all of the publicly indexable Web in the near future. However, is it predicted that the cost of indexing and storage will decline over time

S. Kirkpatrick, Jr. C. D. Gelatt, and M. Vecchi, *Optimization by simulated annealing*, Science 220(4598) (1983), 671-680.

This paper is cited in the following contexts:

M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 177 - Appeared: IEEE ICASSP, San Francisco, March 1992, vol. III, pp.45-48. - **GIBBS RANDOM FIELDS: - TEMPERATURE AND PARAMETER ANALYSIS** - Rosalind W. Picard - M.I.T. Media Lab, E15-392; 20 Ames Street; Cambridge, MA 02139 - picard@media.mit.edu [Details] [Full Text] [Related Articles] [ftp://whitechapel.media.mit.edu/pub/tech-reports/TR-177.ps.Z]

.....**Simulated annealing** is a popular nonlinear optimization technique where a cost function is substituted for E(x), and consequently minimized. There is a key observation in the **simulated annealing** literature that prompts the study of temperature presented in this paper. **Kirkpatrick, et al.*[3]* observed that "more optimization" occurs at certain temperatures than at others.** These favored temperatures are analogous to the physical idea of a "critical temperature," a point that marks transition between different "phases" of the data. The reason for considering these physical.....

.....region, we have shown in earlier work that a similar kind of point, which we call a "transition" temperature, T, does occur [8]. **By measuring the specific heat of the binary process it can be shown to correspond to the same region where the "most optimization" occurs in simulated annealing *[3]*.** For GRF analysis, this region is where the energy fluctuation peaks, and where small changes in the parameters become more significant. In [8] the transition temperature for n = 2 was estimated to be at 1=T = 1:7. This analysis suggests that attempts to estimate parameters should take.....

[3] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. *Optimization by **simulated annealing***. Science, 220(4598):671-680, 1983.

Technical Report No. 9805, Department of Statistics, University of Toronto - **Annealed Importance Sampling** - Radford M. Neal - Department of Statistics and Department of Computer Science - University of Toronto, Toronto, Ontario, Canada - http://www.cs.utoronto.ca/radford/ - radford@stat.utoronto.ca - 18 February 1998 [Details] [Full Text] [Related Articles] [ftp://ftp.cs.utoronto.ca/pub/radford/ais.ps.Z]

.....respect to these transitions. Because such a chain will move between modes only rarely, it will take a long time to reach equilibrium, and will exhibit high autocorrelations for functions of the state variables out to long time lags. **The method of simulated annealing was introduced by *Kirkpatrick, Gelatt, and Vecchi (1983)* as a way of handling multiple modes in an optimization context.** It employs a sequence of distributions, with probabilities or probability densities given by p0(x) to 2 pn(x), in which each pj differs only slightly from pj+1. The distribution p0 is the one of interest. The.....

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983) "*Optimization by **simulated annealing***", Science, vol. 220, pp. 671-680.

[...section deleted...]

Figure 2. An example of how an autonomous citation indexing system can show the context of citations to a given paper. The sentences containing the citations are automatically highlighted.

relative to the increase in the size of the indexable Web (Brin and Page, 1998), resulting in favorable scaling properties for centralized text search engines. In the meantime, an increased number of specialized search services may arise that cover specific types of information.

The use of more expensive and better algorithms (e.g. as in Google) will produce improved page rankings. More information retrieval techniques aimed at the large, diverse, low signal/noise ratio database of the Web will be developed. One interesting possibility is the use of machine learning in order to create query

transformations similar to those used in the specific expressive forms technique discussed earlier.

Metasearch techniques, which combine the results of multiple engines, are likely to continue to be useful when searching for hard to find information, or when comprehensive results are desired. The major Web search engines are also likely to continue to focus on performing queries as quickly as possible, and therefore metasearch engines that perform additional client-side processing (e.g. query term context summaries) may become increasingly popular as these products become more powerful, more effectively combine the results of multiple services, address problems with data fusion from different sources, and learn to deal better with the constantly evolving search services. Improvements in bandwidth should improve the feasibility of metasearch techniques.

Digital libraries incorporating autonomous citation indexing should become more widely available, bringing the benefits of citation indexing to groups which can not afford the commercial services, and improving the dissemination and retrieval of scientific literature.

5 Summary

The Web is revolutionizing information access, however current techniques for access to both general and scientific information on the Web provide room for much improvement. The Web search engines are limited in terms of coverage, recency, how well they rank query results, and the query options they support. Access to the growing body of scientific literature on the publicly indexable Web is limited by the lack of organization and because the major search engines do not index Postscript or PDF documents. We have discussed several fruitful research directions that will improve access to general and scientific information, and greatly enhance the utility of the Web: improved ranking methods, metasearch engines, softbots, and autonomous citation indexing. It is not clear how availability will evolve, because this depends on how the Web emerges as a business platform for publishers. Nevertheless improved ways to do basic searching, and specialized citation searching are likely to evolve and replace present methods, and will greatly increase the utility of the Web over what is available today.

Acknowledgments

We thank H. Stone and the reviewers for very useful comments and suggestions.

References

Barrie, J. and Presti, D. (1996), ‘The World Wide Web as an instructional tool’, *Science* **274**, 371–372.

Boyan, J., Freitag, D. and Joachims, T. (1996), A machine learning architecture for optimizing Web search engines, *in* 'Proceedings of the AAAI Workshop on Internet-Based Information Systems'.

Brin, S. and Page, L. (1998), The anatomy of a large-scale hypertextual Web search engine, *in* 'Seventh International World Wide Web Conference', Brisbane, Australia.

Cameron, R. D. (1997), 'A universal citation database as a catalyst for reform in scholarly communication', *First Monday* **2**(4), `http://www.firstmonday.dk/issues/issue2_4/cameron/index.html`.

Etzioni, O. and Weld, D. (1994), 'A softbot-based interface to the Internet', *Communications of the ACM* **37**(7), 72–76.

Garfield, E. (1979), *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*, Wiley, New York.

Giles, C. L., Bollacker, K. and Lawrence, S. (1998), CiteSeer: An automatic citation indexing system, *in* I. Witten, R. Akscyn and F. M. Shipman III, eds, 'Digital Libraries 98 - The Third ACM Conference on Digital Libraries', ACM Press, Pittsburgh, PA, pp. 89–98.

Inktomi (1997), '`http://www.inktomi.com/press4.html`'.

Kleinberg, J. (1998), Authoritative sources in a hyperlinked environment, *in* 'Proceedings ACM-SIAM Symposium on Discrete Algorithms', San Francisco, California, pp. 668–677, `http://simon.cs.cornell.edu/home/kleinber/auth.ps`.

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

Lawrence, S. and Giles, C. L. (1998*b*), 'Searching the World Wide Web', *Science* **280**(5360), 98–100.

Selberg, E. and Etzioni, O. (1995), Multi-service search and comparison using the MetaCrawler, *in* 'Proceedings of the 1995 World Wide Web Conference'.

Seltzer, R., Ray, E. and Ray, D. (1997), *The AltaVista Search Revolution: How to Find Anything on the Internet*, McGraw-Hill.

Steinberg, S. (1996), 'Seek and ye shall find (maybe)', *Wired* **4**(5).

Taubes, G. (1996), *Science* **271**, 764.