Accessibility of information on the web

Search engines do not index sites equally, may not index new pages for months, and no engine indexes more than about 16% of the web. As the web becomes a major communications medium, the data on it must be made more accessible.

Steve Lawrence and C. Lee Giles

The publicly indexable World-Wide Web now contains about 800 million pages, encompassing about 6 terabytes of text data on about 3 million servers. The web is increasingly being used in all aspects of society; for example, consumers use search engines to locate and buy goods, or to research many decisions (such as choosing a holiday destination, medical treatment or election vote). Scientists are increasingly using search engines to locate research of interest: some rarely use libraries, locating research articles primarily online; scientific editors use search engines to locate potential reviewers. Web users spend a lot of their time using search engines to locate material on the vast and unorganized web. About 85% of users use search engines to locate information[1], and several search engines consistently rank among the top ten sites accessed on the web[2].

The Internet and the web are transforming society, and the search engines are an important part of this process. Delayed indexing of scientific research might lead to the duplication of work, and the presence and ranking of online stores in search-engine listings can substantially affect economic viability (some websites are reportedly for sale primarily based on the fact that they are indexed by Yahoo).

We previously estimated[3] that the publicly indexable web contained at least 320 million pages in December 1997 (the publicly indexable web excludes pages that are not normally considered for indexing by web search engines, such as pages with authorization requirements, pages excluded from indexing using the robots exclusion standard, and pages hidden behind search forms). We also reported that six major public search engines (AltaVista, Excite, HotBot, Infoseek, Lycos and Northern Light) collectively covered about 60% of the web. The largest coverage of a single engine was about one-third of the estimated total size of the web.

How much information?

We have now obtained and analysed a random sample of servers to investigate the amount and distribution of information on the web. During 2–28 February 1999, we chose random Internet Protocol (IP) addresses, and tested for a web server at the standard port. There are currently 256^4 (about 4.3 billion) possible IP addresses (IPv6, the next version of the IP protocol which is under development, will increase this substantially); some of these are unavailable while some are known to be unassigned. We have tested random IP addresses (with replacement), and have estimated the total number of web servers using the fraction of tests that successfully locate a server. Many sites are temporarily unavailable because of Internet connectivity problems or web-server downtime, and to minimize this effect, we rechecked all IP addresses after a week.

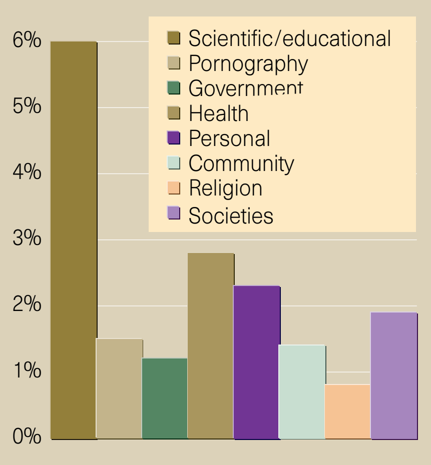

Figure 1 **Distribution of information on the publicly indexable web as of February 1999. About 83% of servers contain commercial content (for example, company home pages). The remaining classifications are shown above. Sites may have multiple classifications.**

Testing 3.6 million IP addresses (requests timed out after 30 seconds of inactivity), we found a web server for one in every 269 requests, leading to an estimate of 16.0 million web servers in total. For comparison, Netcraft found 4.3 million web servers in February 1999 based on testing known host names (aliases for the same site were considered as distinct hosts in the Netcraft survey (www.netcraft.com/survey/)). The estimate of 16.0 million servers is not very useful, because there are many web servers that would not normally be considered part of the publicly indexable web. These include servers with authorization requirements (including firewalls), servers that respond with a default page, those with no content (sites 'coming soon', for example), web-hosting companies that present their home page on many IP addresses, printers, routers, proxies, mail servers, CD-ROM servers and other hardware that provides a web interface. We built a database of regular expressions to identify most of these servers. For the results reported here, we manually classified all servers and removed servers that are not part of the publicly indexable web. Sites that serve the same content on multiple IP addresses were accounted for by considering only one address as part of the publicly indexable web.

Our resulting estimate of the number of servers on the publicly indexable web as of February 1999 is 2.8 million. Note that it is possible for a server to host more than one site. All further analysis presented here uses only those servers considered part of the publicly indexable web.

To estimate the number of indexable web pages, we crawled all the pages on the first 2,500 random web servers. The mean number of pages per server was 289, leading to an estimate of the number of pages on the publicly indexable web of about 800 million. It is important to note that the distribution of pages on web servers is extremely skewed, following a universal power law[4]. Many sites have few pages, and a few sites have vast numbers of pages, which limits the accuracy of the estimate. The true value could be higher because of very rare sites that have millions of pages (for example, GeoCities reportedly has 34 million pages), or because some sites could not be crawled completely because of errors.

The mean size of a page was 18.7 kilobytes (kbytes; median 3.9 kbytes), or 7.3 kbytes (median 0.98 kbytes) after reducing the pages to only the textual content (removing HTML tags, comments and extra white space). This allows an estimate of the amount of data on the publicly indexable web: 15 terabytes (Tbytes) of pages, or 6 Tbytes of textual content. We also estimated 62.8 images per web server and a mean image size of 15.2 kbytes (median 5.5 kbytes), leading to an estimate of 180 million images on the publicly indexable web and a total amount of image data of about 3 Tbytes.

We manually classified the first 2,500 randomly found web servers into the classes shown in Fig. 1. About 6% of web servers have scientific/educational content (defined here as university, college and research lab servers). The web contains a diverse range of scientific material, including scientist, university and project home pages, preprints, technical reports, conference and journal papers, teaching resources, and databases (for example, gene sequences, molecular structures and image libraries). Much of this material is not available in traditional databases. Research articles that do become available in traditional databases are often available earlier on the web[5] (for example, scientists may post preprints on their home pages). The high value of the scientific information on the web, and the relatively small percentage of servers that contain the bulk of that information, suggest that an index of all scientific information on the web would be feasible and very valuable.

We also analysed metadata on the home pages of each server using the HTML META tag. Many different META tags are used, most of which encode details that do not identify the content of the page. We used the existence of one or more of the keywords or description tags as evidence that effort was put into describing the content of the site. We found that 34.2% of servers contain such metadata on their home page. The low usage of the simple HTML metadata standard suggests that acceptance and widespread use of more complex standards, such as XML or Dublin Core[6], may be very slow (0.3% of sites contained metadata using the Dublin Core standard). We also noted great diversity in the HTML META tags found, with 123 distinct tags, suggesting a lack of standardization in usage.

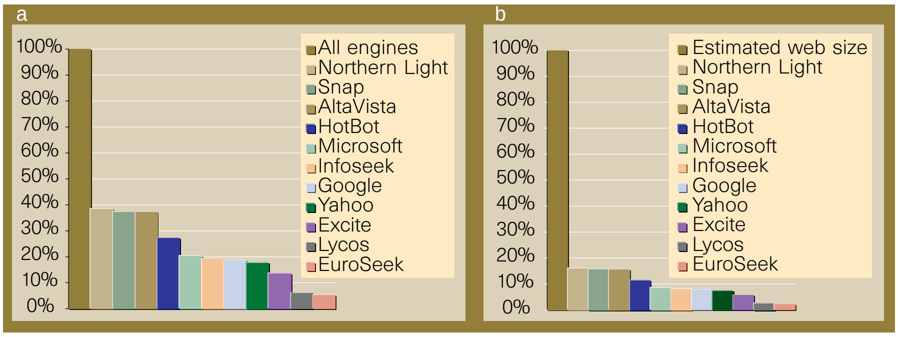

Figure 2 **Relative coverage of the engines for the 1,050 queries used during the period 25–28 February 1999. Note that these estimates are of relative coverage for real queries, which may differ from the number of pages actually indexed because of different indexing and retrieval techniques.** a, **Coverage with respect to the combined coverage of all engines;** b, **coverage with respect to the estimated size of the web.**

Search engine performance

We previously provided coverage results for six major web search engines in December 1997 (ref. 3). We provide updated results here for February 1999 using 11 major full-text search engines (in alphabetical order): AltaVista, EuroSeek, Excite, Google, HotBot, Infoseek, Lycos, Microsoft, Northern Light, Snap and Yahoo. We analysed the Yahoo database provided by Inktomi, not its hierarchical taxonomy (which is much smaller). HotBot, Microsoft, Snap and Yahoo all use Inktomi as a search provider, but the results differ between these engines (due to filtering and/or different underlying Inktomi databases). We did not include Northern Light's 'special collection' (documents that are not part of the publicly indexable web).

We performed our study by analysing the responses of the search engines to real-world queries, as we are interested in the relative coverage of the engines for real queries, which can be substantially different from the relative number of pages indexed by each engine. The engines differ in the specific details of indexing and retrieval, for example there may be different maximum limits on the size of pages indexed by each engine, restrictions on which words are indexed or limits on processing time used per query. This means that comparison has to be done on real queries; comparing the size of the engine databases (readily available from the engines) would lead to misleading results.

In our previous study[3], we used 575 queries originally performed by the employees of NEC Research Institute (mostly scientists). We retrieved the entire list of documents for each query and engine, and then retrieved every individual document for analysis. Only documents containing all of the query terms are included because the search engines can return documents that do not contain the query terms (for example, documents with morphological variants or related terms). These documents may be relevant to the query, but they prevent accurate comparison across search engines. In this study, we have increased the number of search engines analysed to 11, expanded the number of queries to 1,050, and transformed queries to the advanced syntax for AltaVista which allows retrieval of more than 200 results.

We used automatic and manual consistency checks to ensure that responses from the search engines were being correctly processed. The relative coverage of each engine is shown in Table 1 and Fig. 2. Note that Lycos and HotBot returned only one page per site (for Infoseek we used the option not to group results by site). Northern Light has increased substantially in coverage since December 1997, and is now the engine with the greatest coverage (for our queries).

To express the coverage of the engines with respect to the size of the web, we need an absolute value for the number of pages indexed by one of the engines. We used 128 million pages for Northern Light, as reported by the engine itself at the time of our experiments. Our results show that the search engines are increasingly falling behind in their efforts to index the web (Fig. 2).

The overlap between the engines remains relatively low; combining the results of multiple engines greatly improves coverage of the web during searches. The estimated combined coverage of the engines used in the study is 335 million pages, or 42% of the estimated total number of pages. Hence, a substantial improvement in web coverage can be obtained using metasearch engines, such as MetaCrawler[7], which combine the results of multiple engines.

When searching for 'not x', where 'x' is a term that does not appear in the index, AltaVista (Boolean advanced search with the 'count matching pages' option) and Northern Light report the total number of pages in their index. Figure 3 shows the reported number of pages indexed by these engines between 1 November 1998 and 1 April 1999.

Table 1 Statistics for search-engine coverage and recency

Search engine	*Northern Light*	*Snap*	*AltaVista*	*HotBot*	*Microsoft*	*Infoseek*	*Google*	*Yahoo*	*Excite*	*Lycos*	*EuroSeek*	*Average*
Coverage with respect to combined coverage (%)	38.3	37.1	37.1	27.1	20.3	19.2	18.6	17.6	13.5	5.9	5.2	n/a
95% confidence interval	±0.82	±0.75	±0.77	±0.64	±0.51	±0.82	±0.69	±0.61	±0.46	±0.30	±0.40	n/a
Coverage with respect to estimated web size	16.0	15.5	15.5	11.3	8.5	8.0	7.8	7.4	5.6	2.5	2.2	n/a
Percentage of invalid links	9.8	2.8	6.7	2.2	2.6	5.5	7.0	2.9	2.7	14.0	2.6	5.3
Mean age of new matching documents (days)	141	240	166	192	194	148	n/a	235	206	174	n/a	186
Median age of new matching documents (days)	84	91	33	51	57	60	n/a	76	47	174	n/a	57

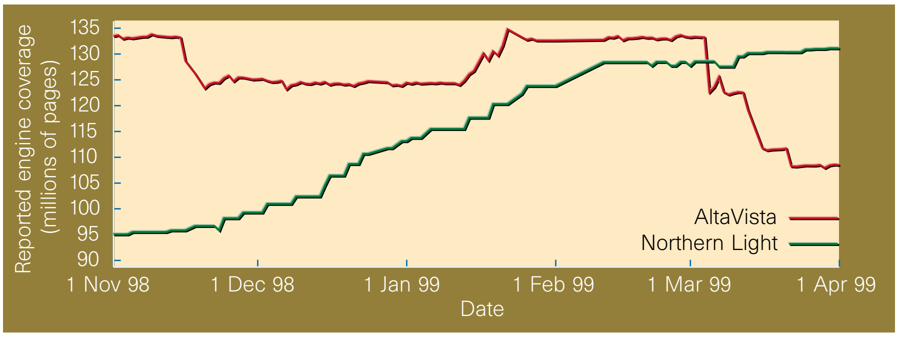

Figure 3 **Number of pages indexed by AltaVista and Northern Light between 1 November 1998 and 1 April 1999 (sampled daily), as reported by the engines themselves.**

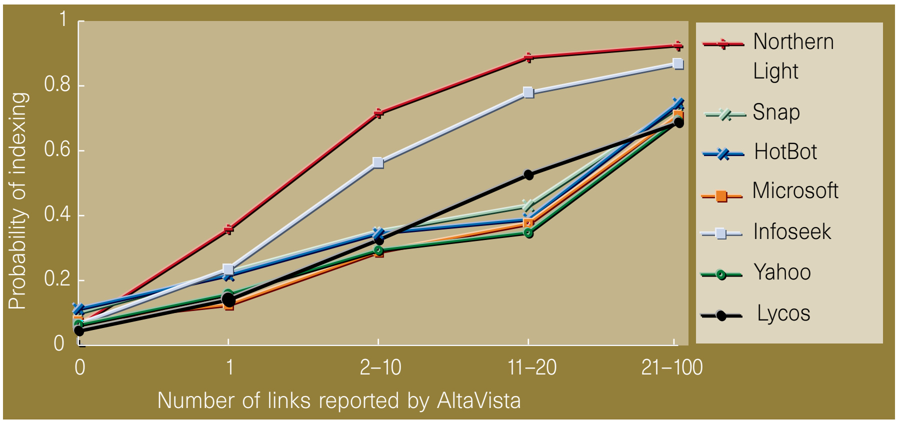

Figure 4 **Probability of indexing random sites versus 'popularity' or connectedness of sites. Sites with few links to them have a low probability of being indexed.**

The number of reported pages indexed by AltaVista has decreased significantly at times, suggesting possible difficulty in maintaining an index of more than 100 million pages.

We next looked at statistics that indicate how up to date the search-engine databases are. Table 1 shows the results of two investigations for each search engine. The first investigation looks at the percentage of documents reported by each engine that are no longer valid (because the page has moved or no longer exists). Pages that timed out are not included. The indexing patterns of the search engines vary over time, and the engine with the most up-to-date database may not be the most comprehensive engine.

We have also looked at the age of new documents found that match given queries by tracking the results of queries using an internal search engine. Queries were repeated at the search engines daily, and new pages matching the queries were reported to the searchers. We monitored the queries being tracked, and when new documents were found, we logged the time since the documents were last modified (when available), which provides a lower bound for the time span between when documents matching a query are added or modified, and the time that they are indexed by one of the search engines (new queries are not included in the study for the first week). The mean age of a matching page when indexed by the first of the nine engines is 186 days, while the median age is 57 days (these averages are over all documents, not over search engines) (Table 1). Although our results apply only to pages matching the queries performed and not to general web pages, they do provide evidence that indexing of new or modified pages can take several months or longer.

Why do search engines index such a small fraction of the web? There may be a point beyond which it is not economical for them to improve their coverage or timeliness. The engines may be limited by the scalability of their indexing and retrieval technology, or by network bandwidth. Larger indexes mean greater hardware and maintenance expenses, and more processing time for at least some queries on retrieval. Therefore, larger indexes cost more to create, and slow the response time on average. There are diminishing returns to indexing all of the web, because most queries made to the search engines can be satisfied with a relatively small database. Search engines might generate more revenue by putting resources into other areas (for example, free e-mail). Also, it is not necessary for a page to be indexed for it to be accessible: a suitable query might locate another page that links to the desired page.

Not equal access

One of the great promises of the web is that of equalizing the accessibility of information. Do the search engines provide equal access? They typically have two main methods for locating pages: following links to find new pages; and user registration. Because the engines follow links to find new pages, they are more likely to find and index pages that have more links to them. The engines may also use various techniques for choosing which pages to index. Therefore, the engines typically index a biased sample of the web.

To analyse bias in indexing, we measured the probability of a site being indexed as a function of the number of links to the site. We used AltaVista to obtain an estimate of the number of links to the home pages of the 2,500 random web servers found earlier. We then tested the existence of the sites in all the other engines that support URL search. Figure 4 shows the probability of each engine indexing a random site versus the number of links to the site. A very strong trend can be seen across all engines, where sites with few links to them have a low probability of being indexed and sites with many links to them have a high probability of being indexed. Considering the relative coverage results, Infoseek has a higher probability of indexing random sites than might be expected. This indicates that Infoseek may index the web more broadly (that is, index more sites instead of more pages per site).

Not only are the engines indexing a biased sample of the web, but new search techniques are further biasing the accessibility of information on the web. For example, the search engines DirectHit and Google make extensive use of measures of 'popularity' to rank the relevance of pages (Google uses the number and source of incoming links while DirectHit uses the number of times links have been selected in previous searches). The increasing use of features other than those directly describing the content of pages further biases the accessibility of information. For ranking based on popularity, we can see a trend where popular pages become more popular, while new, unlinked pages have an increasingly difficult time becoming visible in search-engine listings. This may delay or even prevent the widespread visibility of new high-quality information. □

Steve Lawrence and C. Lee Giles are at NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540, USA.
e-mail: lawrence@research.nj.nec.com

1. Graphic, Visualization, and Usability Center. *GVU's 10th WWW User Survey* (October 1998); http://www.gvu.gatech.edu/user_surveys/survey-1998-10/
2. Media Metrix (January 1999); http://www.mediametrix.com/PressRoom/Press_Releases/02_22_99.html
3. Lawrence, S. & Giles, C. L. *Science* **280,** 98–100 (1998).
4. Huberman, B. A. & Adamic, L. A. *Evolutionary Dynamics of the World Wide Web* (1999); http://www.parc.xerox.com/istl/groups/iea/www/growth.html
5. Lawrence, S., Giles, C. L. & Bollacker, K. *IEEE Computer* **32**(6) (1999).
6. Dublin Core. *The Dublin Core: A Simple Content Description Model for Electronic Resources* (1999); http://purl.oclc.org/dc/
7. Selberg, E. & Etzioni, O. *IEEE Expert* 11–14, Jan.–Feb. (1997).