Searching the World Wide Web

Steve Lawrence and C. Lee Giles*

The coverage and recency of the major World Wide Web search engines was analyzed, yielding some surprising results. The coverage of any one engine is significantly limited: No single engine indexes more than about one-third of the "indexable Web," the coverage of the six engines investigated varies by an order of magnitude, and combining the results of the six engines yields about 3.5 times as many documents on average as compared with the results from only one engine. Analysis of the overlap between pairs of engines gives an estimated lower bound on the size of the indexable Web of 320 million pages.

The Internet has grown rapidly since its inception in December 1969 (*1*) and is anticipated to expand 1000% over the next few years (*2*). The amount of scientific information and the number of electronic journals on the Internet continue to increase [about 1000 journals as of 1996 (*2*, *3*)]. The Internet and the World Wide Web (the Web) represent significant advancements for the retrieval and dissemination of scientific and other literature and for the advancement of education (*2*, *4*). With the introduction of full-text search engines such as AltaVista (www.altavista.digital.com), Excite (www.excite.com), HotBot (www.hotbot.com), In- foseek (www.infoseek.com), Lycos (www.lycos.com), and Northern Light (www.nlsearch.com), the Web can be viewed as a searchable 15-billion-word encyclopedia (*2*). Immediate access to all scientific literature has long been a dream of scientists (*5*), and the Web search engines have made a large and growing body of scientific literature and other information resources accessible within seconds. Scientific information retrieval and literature search, previously dominated by librarians, is now directly available to a widespread group of scientists (*5*).

The major search engine companies have often claimed that they can keep up with the size of the Web [for example, see (*6*)], that is, that they can continue to index close to the entire Web as it grows. However, the Web is a distributed, dynamic, and rapidly growing (*7*) information resource, which presents difficulties for traditional information retrieval technologies. Traditional information retrieval systems were designed for different environments and have typically been used for indexing a static collection of directly accessible documents (*8*). The nature of the Web brings up important questions as to whether the centralized architecture of the search engines can keep up with the expanding number of documents, and if they can regularly update their databases to detect modified, deleted, and relocated information. The answers to these questions impact on the best search methodology to use when searching the Web and on the future of Web search technology.

A number of comparisons have provided relative coverage information for the Web search engines. Typically, these tests involve running a set of queries on a number of engines and reporting the number of results returned by each engine. Results of these comparisons are of limited value because search engines can (and often do) return documents that do not contain the query terms. This behavior can occur because (i) the information retrieval technology used by the engine may not require an exact match (for example, Excite uses "concept-based clustering," and Infoseek uses morphology; these engines can return documents with related words), (ii) documents may no longer exist (an engine that never deletes invalid documents would be at an advantage), and (iii) documents may still exist but may have changed and no longer contain the query terms. Although the additional documents may be relevant to the query, they prevent accurate estimation of the coverage of each engine on the basis of the reported number of results.

Selberg and Etzioni (*9*) presented results based on the usage logs of their MetaCrawler meta search service in 1995 (because of substantial changes in the search engines and the Web, their results would be significantly different if repeated now). Their results are informative but limited. They present the "market share" of each engine, which is the percentage of documents that users follow that originated from each of the search engines. These results are limited for a number of reasons, including the fact that (i) relevance is difficult to determine without viewing the pages, and (ii) presentation order affects user relevance judgments (*10*). They also present results on the percentage of unique documents returned and the coverage of each engine. Their results suggest that each engine covers only a fraction of the Web, but this conclusion cannot be made from their experiments because they only considered the percentage of unique documents out of the top few documents returned by each engine. The search engines return documents in different orders, and Selberg and Etzioni did not distinguish between the following: The engines may cover only a fraction of the Web, or they may cover the entire Web but return different documents among the first few documents listed, because the query results are ranked differently by different engines.

We have produced statistics on the coverage of the major Web search engines, the estimated size of the Web, and the recency of the search engine databases. The following six major full-text search engines were considered (in alphabetical order): AltaVista, Excite, HotBot, Infoseek, Lycos, and Northern Light. A common perception is that these engines index roughly the same documents and that they index a relatively large proportion of the Web.

To compare the number of documents returned by different search engines, we analyzed the search engines' responses to queries performed by employees of the NEC Research Institute (mostly scientists). Our overall methodology was to retrieve the entire list of matching documents from all engines and then retrieve all of the individual documents for analysis. A number of con-

Table 1. Estimated coverage of each engine with respect to the combined coverage of all six (averaged over 575 queries performed during 15 to 17 December 1997), along with the 95% confidence interval (C.I.). HotBot is the most comprehensive in this comparison. 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. Note also that the results may be partly due to different indexing rather than different database sizes: Different engines may not index identical words for the same document (for example, the engines typically impose a maximum file size and effectively truncate oversized documents). However, changes in the results due to different indexing are reflective of the coverage of the engines.

Search engine	Coverage (%)	95% C.I. (%)
HotBot	57.5	±1.3
AltaVista	46.5	±1.3
Northern Light	32.9	±1.1
Excite	23.1	±0.86
Infoseek	16.5	±1.0
Lycos	4.41	±0.42

Computer Science, NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, USA. E-mail: lawrence@research.nj.nec.com (S.L.) or giles@research.nj.nec.com (C.L.G.)

*Also with the Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

straints were imposed. First, the entire list of documents matching the query must have been retrieved from all of the search engines in order for a query to be included in the study. This constraint is important because, as mentioned before, the order in which the engines rank documents varies between engines. Consider a query that results in more than 1000 documents from each engine. If only the first 200 documents from each engine are compared, then many unique URLs (uniform resource locators) may be found. However, we would not be able to determine if the engines were indexing unique URLs or if they were indexing the same URLs but returning different subsets of these URLs in the first 200 documents. Second, for all of the documents that each engine lists as matching the query, we attempted to download the full text of the corresponding URL. Only documents that could be downloaded and actually contained the query terms were counted. This constraint is important because, as detailed above, the search engines can and do return documents that do not contain the query terms.

There were a number of other important details about the analysis. Duplicates were removed when considering the total number of documents returned by one engine or by a combination of engines, including identical pages with different URLs (*11*). Only lowercase queries were considered because different engines treat capitalized queries differently (for example, AltaVista returns only capitalized results for capitalized queries). An individual page time-out of 60 seconds was used; pages that timed out were not included in the analysis. A fixed maximum of 600 documents per query was used (from all engines combined after the removal of duplicates); queries returning more documents were not included (*12*). Only documents that contained the exact query terms were counted. For example, the word "crystals" in a document would not match a query term of "crystal"; the nonplural form of the word would have to exist in the document in order for the document to be counted as matching the query. (This constraint was necessary because different engines use different morphology rules.) Queries with special characters or common "stop" words such as "the" were not used, because the various engines treat special characters differently and use different stop words. HotBot and AltaVista list alternate pages in a special format; these pages were included in the statistics (as they were for the engines that do not specifically identify alternate pages). The "special collection" of Northern Light (premier documents that are not part of the publicly indexable Web) was not used.

We analyzed 575 queries that satisfied these constraints (Fig. 1 and Table 1). The queries were performed during 15 to 17 December 1997 and were taken from queries initially made by NEC employees in the course of their normal work (during a period of about 3 months before the experiments). We manually checked that all results were retrieved from each engine and were parsed correctly because the engines periodically change their formats for listing documents and for requesting the next page of documents (a number of consistency checks were also used to detect temporary failures and changes in the search engine response formats).

We estimated the size of the Web on the basis of an analysis of the overlap among the engines. There are a number of important biases that should be considered. Search engines typically do not consider indexing documents that are hidden behind search forms and are excluded from some documents by the "robots exclusion standard" or authentication requirements. Therefore, we expect the true size of the Web to be much larger than estimated here. However, search engines are unlikely to start indexing these documents in the near future, and it is therefore of interest to estimate the size of the Web that the engines do consider indexing (hereafter referred to as the "indexable Web"). Accurate estimation of the size of the Web based on the overlap among the engines is difficult, because we assume that the engines do not sample the Web independently when they choose pages to index. Each search engine allows users to register their pages with the engine, and it is reasonable to assume that many users will register their pages at several of the engines. Therefore, the pages indexed by each engine will be partially dependent. A second source of statistical dependence between the sampling performed by each engine comes from the fact that search engines are typically biased toward indexing pages that are linked to other pages, that is, more popular pages. With this in mind, we estimate the size of the Web using combinations of two engines (Fig. 2), from the smallest two to the largest two (in terms of our coverage results). It is reasonable to expect that larger engines will have lower dependence because they can index more pages other than the pages that users register and they can index more of the less popular

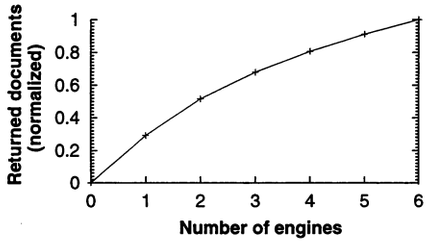

Fig. 1. Coverage as the number of search engines is increased (averaged over 575 queries performed during 15 to 17 December 1997; all results are normalized to the value for six engines). For one to five engines, the average is over all combinations of the engines, which is averaged for each query and then averaged over queries. Significantly more documents are returned as the number of search engines is increased.

Table 2. Estimated size of the portion of the Web that can be indexed from analysis of the overlap between pairs of engines, from the smallest two to the largest two (in terms of our coverage results). The analysis is limited to the 302 queries returning ≥50 documents (to avoid difficulty when $n_0 = 0$), and for each pair of engines *a* and *b*, the estimates derived from p_a and p_b were averaged. Note that the estimate from the smallest two engines is smaller than the actual combined coverage of the six engines used in the study (about 190 million pages). We conclude that this is a result of the statistical dependence between the sampling of the individual engines.

Search engines	Indexable Web (millions of pages)	95% C.I.
Lycos and Infoseek	90	± 6
Infoseek and Excite	220	±16
Excite and Northern Light	230	±15
Northern Light and AltaVista	230	±13
AltaVista and HotBot	320	±34

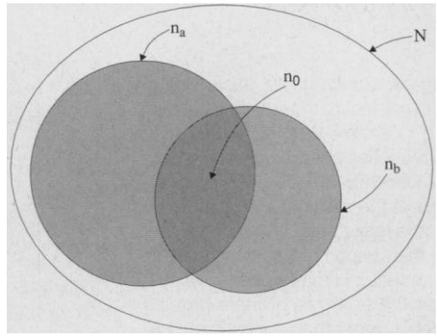

Fig. 2. In order to estimate the size of the indexable Web (*N*, the total number of documents on the Web excluding pages not considered by the search engines), the overlap between pairs of engines was analyzed. Consider the overlap between two engines *a* and *b*. Assuming that each engine samples the Web independently, the quantity n_0/n_b, where n_0 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 as s_a/p_a, where s_a is the number of pages indexed by engine *a*. At the time of the tests, HotBot had reportedly indexed 110 million pages (*16*). We used the relative coverage values as in Table 1 to estimate the number of pages indexed by the other engines.

pages on the Web. Therefore, we expect that the estimated size of the Web, using an assumption of independence, will be more accurate as the engine sizes increase. Indeed, the estimated size of the Web tends to increase when considering the overlap between the larger engines (Table 2).

Using the estimate that the indexable Web contains 320 million pages [from the overlap between the largest two engines (Table 2)], we can express the engine coverage estimates in terms of the fraction of the indexable Web that the individual engines cover: HotBot, 34%; AltaVista, 28%; Northern Light, 20%; Excite, 14%; Infoseek, 10%; and Lycos, 3% (Fig. 3).

Currently available estimates of the size of the Web vary significantly. The Internet Archive uses an estimate of 80 million pages (excluding non-text items such as images and sounds) (*13*). Forrester Research estimates that there are more than 75 million pages (*14*). AltaVista's chief technical officer, Louis Monier, now estimates that the Web contains 100 to 150 million pages (*15*). Wired Digital reports that the Web contained about 175 million pages as of December 1997 (*16*). Tom Mitchell extrapolated from sizes reported in the literature in 1995 and 1996 to produce a current estimate of 200 million pages (*17*). On the basis of our results, it appears that existing estimates significantly underestimate the size of the Web.

We also investigated the percentage of documents reported by each engine that are no longer valid (because the page has moved or no longer exists) and the median age of the documents returned by each engine. These investigations provide some information on the recency of the search engine databases. For the experiments run on 15 to 17 December 1997, the percentages of invalid links were, from best to worst, 1.6% for Lycos, 2.0% for Excite, 2.5% for AltaVista, 2.6% for Infoseek, 5.0% for Northern Light, and 5.3% for HotBot (pages that timed out were not included in these statistics). In comparison with the results of similar experiments performed in August 1997, the ranking of the engines in terms of the percentage of invalid links has changed significantly. Analysis of the median age of documents returned by the engines showed similar changes from the experiments performed in August 1997. Our results suggest that the indexing patterns of the engines vary significantly over time, and that the engine with the most recent pages may not be the most comprehensive engine (one factor involved here may be a tradeoff between the database size and update frequency).

A number of conclusions can be drawn from these experiments. The coverages of the search engines investigated vary by an order of magnitude. An estimated lower bound on the size of the indexable Web is 320 million pages. The engines index only a fraction of the total number of documents on the Web; the coverage of any one engine is significantly limited. On the basis of our estimate of the size of the indexable Web, the individual engines cover from 3 to 34% of the indexable Web. The engines may be limited by network bandwidth, disk storage, computational power, or a combination of these items [despite claims to the contrary (*6*)]. Combining the results of multiple engines can significantly increase coverage: Combining the six engines in this study covered about 3.5 times as much of the Web as one engine. If only two engines are used, the two engines with the largest coverage are currently HotBot and AltaVista.

Scientists often search for information that does not occur in many places on the Web (for example, the home page of another scientist or information about a specific paper may not be duplicated or have many links referring to it). Given that the coverage of any one search engine is limited, the simplest means of improving the coverage of Web search engines is to combine the results of multiple engines, as is done with meta search engines such as MetaCrawler (www.metacrawler.com). Another alternative is to combine available information sources such as the major search engines with automated online searching. One example is the Internet "softbot" (*18*). The softbot transforms queries into goals and uses a planning algorithm to generate a sequence of actions in order to satisfy the goal. The planner has extensive knowledge of the information sources that it accesses. One successful softbot is the AHOY! service, which locates home pages for individuals (*19*). In a study where Shakes *et al.* searched for the home pages of 582 researchers, AHOY! was able to locate more home pages than Meta-Crawler (which located more home pages than HotBot or AltaVista) with greatly improved precision. Another possibility for improved searching for scientists is the creation of a search engine designed to keep up-to-date indexes of pages that are important to scientists.

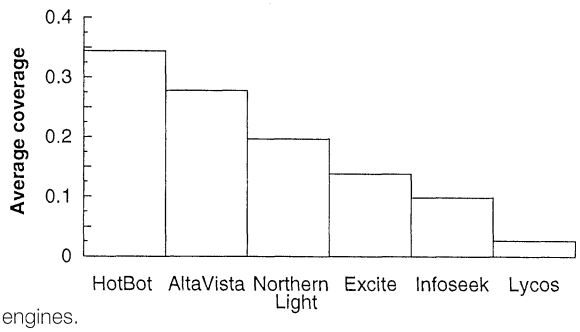

Fig. 3. Coverage of each engine with respect to the estimated size of the indexable Web (averaged over 575 queries performed during 15 to 17 December 1997). The percentage of the indexable Web indexed by the major search engines is lower than is commonly believed. We note that it is reasonable to expect that the true size of the indexable Web is larger than our estimate because of the statistical dependence that remains between the two largest engines.

REFERENCES AND NOTES

1. W. Howe, "When did the Internet start?: A brief capsule history," www.delphi.com/navnet/faq/history.html (9 May 1996); "Network Wizards Internet Domain Survey," www.nw.com/zone/WWW/top.html (July 1997); A. M. Rutkowski, "Internet trends," www.genmagic.com/Internet/Trends/(February 1997).
2. J. M. Barrie and D. E. Presti, *Science* **274**, 371 (1996).
3. G. Taubes, *ibid.* **271**, 764 (1996).
4. P. T. Fox and J. L. Lancaster, *ibid.* **266**, 994 (1994).
5. B. R. Schatz, *ibid.* **275**, 327 (1997).
6. S. G. Steinberg, *Wired* **4** (no. 5), 108 (1996).
7. M. Gray, "Measuring the growth of the Web: June 1993 to June 1995," www.mit.edu/people/mkgray/growth/ (1996).
8. G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, 1983).
9. E. Selberg and O. Etzioni, in *Proceedings of the Fourth International World Wide Web Conference*, Boston, MA, 11 to 14 December 1995, p. 195.
10. M. Eisenberg and C. Barry, *Proceedings of the 49th Annual Meeting of the American Society for Information Science* (Learned Information, Medford, NJ, 1986), p. 80.
11. The URLs were normalized by (i) removing any "index.html" suffix or trailing "/", (ii) removing a port 80 designation (the default), (iii) removing the first segment of the domain name for URLs with a directory depth greater than 1 (to account for machine aliases), and (iv) unescaping any "escaped" characters (for example, %7E in a URL is equivalent to the tilde character).
12. The search engines typically impose an upper limit on the number of documents that can be retrieved [current limits are 200 for AltaVista (simple search), 500 for Infoseek, 1000 for HotBot, Excite, and Lycos, and >10,000 for Northern Light], and we checked to ensure that these limits were not exceeded.
13. M. Cunningham, "Brewster's millions," www.irish-times.com/irish-times/paper/1997/0127/cmp1.html (27 January 1997).
14. C. Guglielmo, *Upside* (November 1997), p. 48.
15. D. Brake, *New Sci.* **154**, 12 (1997).
16. PR Newswire, "Wired Digital's HotBot search site unveils largest Web index; 110 million page database extends and enhances Web power searching" (10 December 1997).
17. T. Mitchell, personal communication.
18. O. Etzioni and D. Weld, *Commun. ACM* **37** (no. 7), 72 (1994).
19. J. Shakes, M. Langheinrich, O. Etzioni, in *Proceedings of the Sixth International World Wide Web Conference*, Santa Clara, CA, 7 to 11 April 1997.

20. We would like to thank A. Grove, B. Horne, B. Krovetz, J. Oliensis, S. Omohundro, H. Stone, L. Williams, P. Yianilos, and the anonymous reviewers for useful comments and suggestions. All registered and unregistered trademarks are the property of their respective owners.

17 April 1997; accepted 10 February 1998