

A technique for measuring the relative size and overlap of public Web search engines

[Krishna Bharat](#) and [Andrei Broder](#)

[DIGITAL, Systems Research Center,](#)
130 Lytton Avenue, Palo Alto, CA 94301, U.S.A.
bharat@pa.dec.com and broder@pa.dec.com

Abstract

Search engines are among the most useful and popular services on the Web. Users are eager to know how they compare. Which one has the largest coverage? Have they indexed the same portion of the Web? How many pages are out there? Although these questions have been debated in the popular and technical press, no objective evaluation methodology has been proposed and few clear answers have emerged. In this paper we describe a standardized, statistical way of measuring search engine coverage and overlap through random queries. Our technique does not require privileged access to any database. It can be implemented by third-party evaluators using only public query interfaces. We present results from our experiments showing size and overlap estimates for HotBot, AltaVista, Excite, and Infoseek as percentages of their total joint coverage in mid 1997 and in November 1997. Our method does not provide absolute values. However using data from other sources we estimate that as of November 1997 the number of pages indexed by HotBot, AltaVista, Excite, and Infoseek were respectively roughly 77M, 100M, 32M, and 17M and the joint total coverage was 160 million pages. We further conjecture that the size of the static, public Web as of November was over 200 million pages. The most startling finding is that the overlap is very small: less than 1.4% of the total coverage, or about 2.2 million pages were indexed by all four engines.

Keywords

Search engines; Coverage; Web page sampling

1. Introduction

Search engine such as [AltaVista](#), [Excite](#), [HotBot](#), [Infoseek](#), and [Lycos](#), are among the most useful and popular services on the Web. Naturally they are the subject of intense scrutiny both in the popular press (see for instance articles in *PC Computing* and *Washington Post*) and within the scientific community (see for instance the proceedings of ASIS 96, WWW6, etc). An excellent [bibliography](#) maintained by Traugott Koch lists over 60 publications on the subject of search service comparisons. A comprehensive Web site dedicated to search engine comparisons at <http://www.searchenginewatch.com/> maintained by Danny Sullivan.

Clearly no search engine can index the entire Web. (A good discussion of this issue appears in [1].) An often debated topic is how much coverage they provide. Typical questions are "Which one has the largest coverage?", "Do they all cover the same pages and differ only in their ranking, or do they have low overlap?", "How many pages are out there and how many are indexed?". These questions are of scientific and public interest but few objective direct evaluation methodologies have been proposed.

The most straightforward way to get answers would be to obtain the list of URLs within each engine's index, and compute the sizes and intersections of these lists. Given the intense competition among

commercial search services, and the fact that such a list is highly prized proprietary information, this certainly will not happen. Furthermore even if the services were to supply a third party with a list of URLs there is no guarantee that they represent legitimate URLs and/or that the corresponding pages were actually fetched and still in the index.

Some published approaches to estimating coverage are based on the number of hits for certain queries as reported by the services themselves. A better variation, used by [Melee's Indexing Coverage Analysis](#) (MICA) is based on the reported count of pages within a particular domain. (MICA uses other refinements, but is still ultimately dependent on the accuracy of the reported numbers.) Unfortunately self-reported counts are not necessarily reliable or comparable. Even assuming no deliberate over-estimation there is no guarantee that the reported counts do not include duplicate documents, aliased URLs, or documents no longer in existence. Furthermore this method does not allow any determination of overlap.

A different approach, used for example by the [Search Engine Watch](#) for their [Search Engine EKG](#), is to observe the access logs of selected sites and determine how much of these sites is visited by various search engine robots. Again there is no guarantee that if an engine has fetched a page, the page will eventually be indexed. Search engines tend to employ various filtering policies to build an effective index.

In an attempt to resolve the questions above we have developed a standardized, statistical way of measuring search engine coverage and overlap through random queries. Our technique does not require privileged access to any database, and can be implemented by independent third-party evaluators with fairly modest computational resources. (For four engines, we could determine their relative size and the relative size of their intersection within a few days. Most of the time is spent waiting for pages to be fetched.) Our method is subject to certain statistical biases that we discuss, but since they tend to favour content-rich pages, this can be viewed as an asset rather than a flaw.

We implemented our measurement technique as described in this paper and did two sets of experiments involving over 10,000 queries each. These show that as of mid 1997, the approximate sizes of HotBot, AltaVista, Excite, and Infoseek, expressed relative to their joint total coverage at that time, were respectively 47%, 39%, 32%, and 18%; and as of November 1997, they were 48%, 62%, 20%, and 17%. Our method does not provide absolute values. However using data from other sources we estimate that as of November the number of pages indexed by HotBot, AltaVista, Excite, and Infoseek were respectively roughly 77 million, 100 million, 32 million, and 17 million and the joint total coverage was 160 million pages. We further conjecture that the size of the static, public Web as of November was about 200 million pages. The most startling finding is that the overlap was very small: less than 1.4% of the total coverage, or about 2.2 million pages appeared to be indexed by all four engines.

In the [next section](#) we overview our estimation algorithm. Its detailed implementation is described in [Section 3](#). In [Section 4](#) we discuss sources of bias in our technique and suggest how they may be overcome or even exploited. [Section 5](#) describes in detail the two sets of experiments that we did to estimate the sizes and overlaps of AltaVista, Excite, Hotbot and Infoseek. Finally [Section 6](#) presents some conclusions.

2. Our approach

2.1. Overview

Our scheme allows a third party to measure the relative sizes and the overlaps of search engine indices;

for every pair of engines E_1 and E_2 we can compute:

- Their relative sizes, that is the ratio: $\text{Size}(E_1) : \text{Size}(E_2)$,
- The fraction of E_1 's database indexed by E_2 , expressed as a percentage of the size of E_1 .

More generally, we can compute the fraction of E_1 's database simultaneously indexed by E_2, E_3, \dots . The scheme does not require privileged access to any engine's database; only the ability to make queries.

The idea behind our approach is simple: Consider sets A and B of size 4 and 12 units respectively. Let the size of their intersection $A \cap B$ be 2 (see Fig. 1). Let us suppose that we do not know any of these sizes but can sample uniformly from any set and check membership in any set. If we sample uniformly from A and we will find that about $1/2$ of the samples are in B as well. Hence the size of A is approximately 2 times the size of the intersection, $A \cap B$. Similarly we will find that the size of B is approximately 6 times the size of $A \cap B$. This will lead us to believe that B is about $6/2 = 3$ times the size of A .

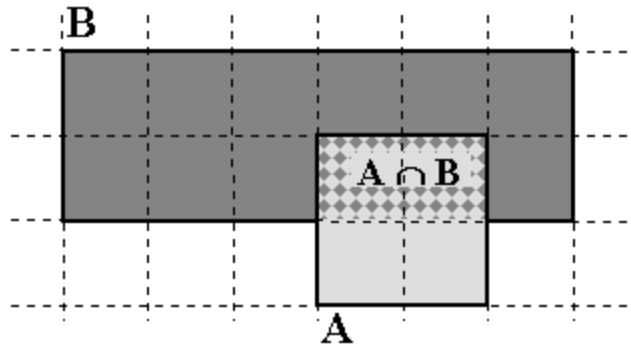


Fig. 1. Computing size ratios from overlap estimates.

More formally, let $\text{Pr}(A)$ represent the probability that an element belongs to the set A and let $\text{Pr}(A \cap B | A)$ represent the conditional probability that an element belongs to both sets given that it belongs to A . Then, $\text{Pr}(A \cap B | A) = \text{Size}(A \cap B) / \text{Size}(A)$ and similarly, $\text{Pr}(A \cap B | B) = \text{Size}(A \cap B) / \text{Size}(B)$, and therefore $\text{Size}(A) / \text{Size}(B) = \text{Pr}(A \cap B | B) / \text{Pr}(A \cap B | A)$.

To implement this idea we need two procedures:

- **Sampling:** A procedure for picking pages uniformly at random from the index of a particular engine.
- **Checking:** A procedure for determining whether a particular page is indexed by a particular engine.

Neither procedure can be implemented perfectly without privileged access to the search engine databases. However, as we explain below, we can construct good approximations that use only queries made via public interfaces. Given these procedures, we estimate overlaps and relative sizes as follows:

- **Overlap estimate:** The fraction of E_1 's database indexed by E_2 is estimated by:

$$\text{Fraction of URLs sampled from } E_1 \text{ found in } E_2.$$

- **Size comparison:** For search engines E_1 and E_2 , the ratio $\text{Size}(E_1)/\text{Size}(E_2)$ is estimated by:

$$\frac{\text{Fraction of URLs sampled from } E_2 \text{ found in } E_1}{\text{Fraction of URLs sampled from } E_1 \text{ found in } E_2}$$

2.2. Alternatives

For any objective evaluation we must check via the public interface whether or not certain test URLs are actually in the indices of the engines being considered. The mere fact that a search engine has encountered a URL or looked at a page (determinable from site access logs) is not sufficient. Hence a checking procedure is inherently necessary for every evaluation scheme.

However in principle sampling can proceed along different lines. For instance, if we had a way to select a page on the Web uniformly at random, we could choose a large set of random pages and for each page test if it is indexed by each of the search engines. This would allow us to estimate the relative sizes and overlaps of the search engines under consideration and also their sizes relative to the entire Web.

Unfortunately choosing pages uniformly at random from the entire Web is practically infeasible: it would require either collecting all valid URLs on the Web, which requires constructing a better Web crawler than any existing one, or the use of sampling methods that do not explore the entire Web. Such methods, based on random walks, have been studied theoretically in other contexts (see e.g. [3] and references therein). They are not easily applicable to the Web, since the Web is a directed graph with highly non-uniform degrees and has many small cuts. Precious little is known about the graph structure of the Web. Hence the length of the random walks required to generate a distribution close to the uniform may be extremely large.

Since sampling at random from the Web directly is not easy, we could use the search engines themselves to generate page samples. It is tempting to use one engine, say E_1 as a random source of URLs to estimate the sizes of two other search engines, say E_2 and E_3 . For a random page from E_1 we could check whether it belongs to E_2 and whether it belongs to E_3 . Given enough samples we could infer the relative sizes of E_2 and E_3 . However such a strategy could work only if the search engine indices are truly independent, and built from random samples of the Web. This is not true in practice. Two engines may well have started at the same starting points, or have taken URL submissions from the same parties, or may use the same policy for prioritizing links and indexing pages. If engine E_1 were used as a source and there is a positive correlation between E_1 's index and E_2 's index, then E_2 will unfairly appear to be bigger than the others. Hence we employ the strategy described in Section 2.1.

3. Implementation

We implement both sampling and checking via queries made to the public interface. Both procedures make use of a precomputed *lexicon* of Web words and their associated frequencies on the Web. This lexicon is derived by analysing a standard corpus of Web pages that is not associated with any indexing engine under consideration. In the process, the expected frequency of words is computed. It is desirable that this corpus be Web based so that (a) words used in Internet parlance are included, and (b) so that their frequencies reflect their pattern of occurrence on the Web at large.

For our experiments we used a broad crawl of roughly 300,000 documents in the [Yahoo!](#) hierarchy to build a lexicon of about 400,000 words. Extremely low frequency words (which may have been typographic errors) were not included in the lexicon. We conjecture that all search engines must use

Yahoo! as one of their starting points and hence the lexicon would be fair to all. Whether indeed the composition and word distribution in our lexicon reflects that of the whole Web is unclear.

In general the choice of a particular lexicon induces a certain bias, as we further discuss below. For instance, such a lexicon may be heavily biased towards English. In principle it is possible to choose a corpus of Web pages in another language and construct the lexicon accordingly. In addition we can impose additional constraints, such as a restriction on the hostname (e.g., *.com) to obtain statistics on a particular section of the Web.

3.1. Query based sampling

We generate a random URL from a search engine by firing a random query at it and selecting a random URL from the result set. In practice it is not possible to fetch the entire result set. Search engines usually do not return more than a few hundred matches and in fact may not even compute the remaining matches. Hence in our experiments we considered only the first 100 results and picked a random page from among them. The URLs listed by a search engine as the best 100 matches depend on the engine's ranking strategy, and this introduces a *ranking bias* in our estimate as we discuss in [Section 4](#).

We experimented with both disjunctive (OR) queries and conjunctive (AND) queries. All queries tend to introduce a *query bias* towards large, content rich pages (see [Section 4](#)). For conjunctive queries one can try to pick keywords such that fewer than 100 results are returned, thus eliminating the ranking bias. Unfortunately this increases the query bias.

Disjunctive (OR) queries are made out of random sets of keywords drawn from the lexicon. We used sets of four. To reduce the query and ranking bias the words need to be chosen so that their frequencies are roughly the same, since some search engines prefer infrequent keywords to rank results.

To compute conjunctive (AND) queries, pairs of random keywords are used (sets of 3 will often result in 0 hits). Keywords are paired carefully to return a small but non-empty set of results. This is done as follows: The list of available keywords is sorted by frequency. A lower and an upper threshold are picked iteratively so that keywords equidistant from the thresholds tend to give between 1 and 100 results in a conjunctive query. Then the keywords contained between the thresholds are randomly sampled. The resulting set of keywords is sorted by frequency, and keywords equidistant from the ends of the list are paired to form queries.

Given a random query in one of the two schemes, a URL is obtained from a designated search engine by selecting a random result page from the top 100 result pages returned by the search engine for the query.

This query based sampling approach is subject to various biases, that is some pages tend to have a higher probability of being selected as a sample. This gives them a higher "weight". Eventually what we end up estimating is the ratio of the total weight of the pages indexed by one engine to the total weight of the pages indexed by another. However the bias introduced seems reasonable or even favourable, in the sense that we tend to give a higher weight to "interesting" pages — namely pages rich in content in our language of choice. This is further discussed in [Section 5](#).

3.2. Query based checking

To test whether a page with a given URL (which was chosen as explained above) has been indexed by a particular search engine, we construct a query meant to strongly identify the page. We call such a query a *strong* query. Ideally the strong query will uniquely identify the page and the result page from a search

engine that has indexed the page should have exactly one result — namely the URL in question. In practice, there may be multiple results. There are several reasons for this:

- The same page may be accessible through multiple aliases.
- Near-identical versions of the page may exist.
- Mirrored copies of the page may be present at other sites.
- In some cases the page may be so devoid of content that a strong query that matches that page will match many other pages as well.

We explain below how we cope with these problems.

The strong query is constructed as follows. The page is fetched from the Web and we analyse the contents to compute a conjunctive query composed of the k (say 8) most significant terms in the page. Significance is taken to be inversely proportional to frequency in the lexicon. Words not found in the lexicon are ignored since they may be typographical errors (which tend to be transient) or words that might not have any discriminative value (e.g., common words in a foreign language.) The strong query is sent to each of the search engines in turn and the results are examined. If one of the result URLs matches the query URL then the query URL is noted as being present in the engine's database.

The matching has three aspects:

(a) Normalization. All URLs are translated to lowercase and optional filenames such as `index.htm[1]`, `home.htm[1]` are eliminated. Relative references of the form `"#..."` and port numbers are removed as well. Hostnames are translated to IP addresses when one of the URLs involves an IP address.

(b) Actual matching. This is done in several ways depending on how strict the matching requirement is:

- **Full URL comparison.** In this case the page is deemed to be present only if a page with the same (normalized) URL is returned by the search engine under consideration.
- **High similarity.** In this case we retrieve all the result pages and compare each in turn with the page we are looking for. (Normalization of URLs is not necessary.) If any page has a similarity above (say) 95% we deem the page present. This compensates for aliasing and mirroring. A fast page similarity technique such that proposed in [2] can be used for this purpose. (We have not yet implemented this approach.)
- **Weak URL comparison.** In this case we do only a hostname comparison. The page is deemed to be present if a page with the same host matches the strong query. This compensates for reorganizations of a particular host but has the risk of being overly generous.
- **Non zero result set.** In this case any URL returned by the strong query is considered a match.

Note that it might turn out that a URL returned by a search engine no longer points to the version that was indexed by the source engine, because the content has changed. We discuss this below.

(c) Filtering out dynamic and low-content pages. Pages that are detected by none of the search engines (including the source) may be regarded as dynamic Web pages and filtered out of the statistics on the assumption that they contain changing content. Similarly, some pages actually have very little textual content which results in an ineffective strong query. (That is, the strong query returns a large number of matches from either the source or the engine being tested.) Ignoring such pages allows us to focus on content rich pages with relatively static content.

The checking step introduces a *checking bias* that we discuss further below. Again our estimates weight

in the favour of pages that are content rich and relatively static. Although this is not representative of the entire Web, it reflects the part of the Web that most users tend to query, and is hence more meaningful as a measure of a search engine's utility. As already explained, query based sampling is inherently biased towards pages rich in text content. Hence, this effect seems inevitable regardless of the checking method used.

4. Bias

In the previous section we have identified several sources of bias. Here we analyse them further and discuss some possible counter-measures. We did not implement these in computing our measurements however.

We classify the biases as follows:

Query bias:

The accessibility of pages via queries tends to vary based on the number of keywords in the document and probability of inclusion in a query. Large, content rich documents tend to have a better chance of matching a query. In general the query bias is determined by our choice of lexicon and/or method of generating queries. Nevertheless, it appears that results obtained using disjunctive queries are fairly similar to results obtained using conjunctive queries (see [Table 1](#), Trial 1 vs. Trial 2), even though the query bias is different. We conjecture that this happens because the proportion of pages having a certain probability under each query type is fairly similar for all sets of pages under consideration. In other words, even though pages are not picked uniformly at random, the number of pages from index A found in index B is proportional to the size of the intersection of A and B divided by the size of A.

Ranking bias:

Search engines introduce a bias by ranking pages. Since only a subset of the pages returned by the query are served up by the search engine the remaining pages are effectively not sampled.

Checking bias:

The method that we use for actual matching and/or our policy with regard to dynamic and low content pages influence the probability that samples are found in the index being tested.

Experimental bias:

We are estimating a moving target. The content of a particular search engine might change during our experiments. Pages are continuously added and deleted and search engines might also decide under load to time-off certain queries and/or make only part of the index available.

Malicious bias:

Under an unlikely scenario, a search engine might rarely or never serve pages that other engines have, thus completely sabotaging our approach.

Statistical sampling error:

We are estimating each measurement from a finite sample. There is a certain probability that the sample average is very different from the value being estimated. This probability can be easily computed from standard statistical formulas. In all cases we used enough samples to keep the 95% confidence suitably narrow. This error can be arbitrarily reduced by using more samples.

Some of these biases can be alleviated, but usually there are trade-offs. To remove the ranking bias we could frame conjunctive queries that return less than, say, 200 documents, although this is hard to guarantee, and retrieve all the documents. Clearly this increases the query bias.

Removing the query bias is more difficult. One could in principle compute $P_{\text{incl}}(u)$, the probability of inclusion of URL u given our process for generating URLs. This depends on the probability of

generation of various queries and the probability of u being matched by each of the queries. If P_{incl} were the same for all URLs then we could select all URLs with equal probability. Since this is not the case (due to the query bias) we can compensate by selecting u with a selection probability of $P_0/P_{incl}(u)$, where P_0 is the minimum value for $P_{incl}(u)$ on the Web. These quantities are very hard to estimate in practice, and since in most cases the selection probability will be very low, this process will take a large number of trials to generate the needed set of URLs.

With privileged access to a search engine it is trivial to generate random URLs from the set indexed by that engine, thus eliminating both ranking and sampling bias. This should show a decrease in overlap, since one would expect that high-content pages which are favoured by both ranking and query bias are likely to be cross-indexed. Preliminary experiments done using privileged access to the AltaVista index confirm this.

5. Measurements

We conducted two sets of measurements: the first set in June and July 1997 (trials 1 and 2), and the second set in November 1997 (trials 3 and 4). Trials 1 and 3 involved 4 term disjunctive queries, and trials 2 and 4 involved 2 term conjunctive queries. In all cases there were approximately 10,000 queries, except for trial 2 which had about 5,000. In each trial the queries were divided among the search engines, so that each query went to one of the search engines. A random URL was selected from the top 100 results as the sample URL, to check for containment in all the search engines. In each case we examined the first page of results (10 results) from the strong query for potential matches. If any search engine returned more than 10 results the test was discarded, since we deemed the strong query to be ineffective.

Engine A (source of URLs)	Engine B (under test)	Trials in June–July 1997						Trials in November 1997					
		Measured overlap (URLs from A also in B)			Inferred relative size (size A/size B)			Measured overlap (URLs from A also in B)			Inferred relative size (size A/size B)		
		1	2	1&2	1	2	1&2	3	4	3&4	3	4	3&4
AltaVista	EX	17%	24%	19%	1.33	1.05	1.21	16%	15%	15%	3.17	3.31	3.23
	HB	37%	38%	37%	0.87	0.77	0.84	39%	38%	39%	1.39	1.17	1.28
	IS	17%	16%	16%	2.03	2.25	2.1	12%	18%	15%	4.08	2.77	3.32
Excite	HB	29%	30%	29%	0.61	0.75	0.65	36%	39%	37%	0.49	0.46	0.47
	IS	11%	10%	11%	1.62	2.38	1.83	14%	15%	14%	1.39	1.26	1.33
	AV	22%	25%	23%	0.75	0.95	0.82	52%	50%	51%	0.32	0.3	0.31
HotBot	IS	12%	12%	12%	2.77	2.91	2.81	16%	13%	15%	2.63	3.21	2.89
	AV	32%	29%	31%	1.15	1.3	1.2	54%	45%	50%	0.72	0.86	0.78
	EX	17%	22%	19%	1.64	1.34	1.53	17%	18%	17%	2.04	2.18	2.11
	AV	34%	36%	31%	0.49	0.44	0.48	50%	50%	50%	0.24	0.36	0.3

Infoseek	EX	19%	23%	20%	0.61	0.42	0.55	19%	19%	19%	0.72	0.79	0.75
	HB	34%	37%	35%	0.36	0.34	0.36	44%	44%	44%	0.38	0.31	0.35

Table 1. Overlap and size estimates (static pages only, full URL comparison)

Statistics were computed for various inclusion criteria (all pages, static pages only) and matching criteria (full URL match, hostname match, any URL). When considering static pages only, the matching criterion determined how many tests were applicable, since, by our definition, at least one search engine needed to contain the page for the test to be applicable. We consistently found that about 10% of the pages sampled were dynamic and another 10% failed to produce a small result set with a strong query. Hence only 80% of the tests were actually considered for the static pages measurements.

5.1. Pairwise overlap and relative size estimates

Table 1 lists the measured overlaps and computed size ratios between search engines for our two sets of trials. In each case we also list statistics for both trials combined. The results shown were computed with full URL matching on static pages. As we discuss in [Section 5.2](#) we felt this to be the most reliable measurement strategy to use. The estimates reflect the state of the world in mid 1997 and in November of 1997. Search engines tend to grow over time and rebuild their indices. *Hence, neither scenario may be representative of the current status.*

The leftmost column lists the source engine for the random pages (Engine A). The second column lists the engine being tested for containing these pages (Engine B). For each trial we list the percentage random pages from A found in B (overlap), and the inferred ratio of sizes: Size(A)/Size(B). To compute the ratio we use the method explained in [Section 2.1](#).

Some variation between the trials in a set is to be expected. A variation in an overlap value will cause a variation in the corresponding size ratio. In most cases the overlap variation appears to be statistically insignificant. The only significant variation in the first set of trials is an apparent increase in Excite's coverage from trial 1 to trial 2. This is probably due to different biases in effect with disjunctive and conjunctive queries. With conjunctive queries there is a query bias towards large, content-rich pages. This might indicate that Excite was selectively indexing content-rich pages at that point in time. The other explanation might be that Excite installed a larger index in the interval between the trials (about 15 days). We observe two mildly significant variations in the second set of trials (i.e., trials 3 and 4): (1) Infoseek indexed more of AltaVista pages accessed with conjunctive queries than disjunctive queries, and (2) Infoseek and AltaVista indexed more of HotBot with disjunctive queries. As before, the first variation might arise because Infoseek selectively indexed content rich pages. The second observation may be due to a ranking bias on HotBot's part towards more "popular" pages by some metric (e.g., content length, in-degree). Such a bias could be more visible with respect to pages fetched with disjunctive queries which generally return large sets. In general, the interplay between ranking biases and page indexing policies of search engines has the potential to produce isolated variations in overlap that are hard to explain. However the overall results seem fairly consistent.

5.2. Comparison of measurement strategies

We considered three URL matching criteria: full URL comparison, hostname matching (weak URL comparison), and considering any URL satisfying the strong query to be a match (non zero result set). In addition to the preferred case of static pages only, we also consider the case when dynamic pages are included.

As mentioned previously 10% of the pages were observed to be dynamic. Hence the overlaps when considering static pages alone were about 10% larger than when all pages were considered. This did not affect size estimates. Quite surprisingly size ratios seemed unaffected by the matching criterion as well. As we weakened the URL matching criterion, the overlap fractions grew correspondingly. Strong queries often list large files such as dictionaries in their result set. Hence, taking a non-zero result set to be a match seems overly generous. While hostname matching may compensate in some cases for URL naming differences due to aliases, we believe full URL matching to be a more dependable scheme, even though it tends to underestimate overlap a little.

5.3. Absolute estimates

The pair-wise size estimates we listed in Table 1 are only approximately consistent. That is the product of two size ratios will not be exactly equal to the corresponding third computed ratio. The values are close enough, however, to rank the engines by size. In June/July 1997 we observed that $Size(HotBot) > Size(AltaVista) > Size(Excite) > Size(Infoseek)$, and in Nov 1997 we observed that $Size(AltaVista) > Size(HotBot) > Size(Excite) > Size(Infoseek)$. The overlaps we observed are rather low. The most startling finding is that less than 1.4% of the total coverage was indexed simultaneously by all four engines.

To reconcile the various pair-wise size ratios and get a full picture we computed best likelihood estimates for engine sizes. Specifically, we computed estimates for the engine sizes so that the sum of squared differences between the resulting estimates for the pairwise overlap (using the experimental ratios) was minimized. Using these estimates for sizes, we then averaged the estimates for the intersections given by the experimental ratios. Finally we normalized our data so that the total coverage of all search engines together equals 100%. The final numbers are presented in [Fig. 2](#).

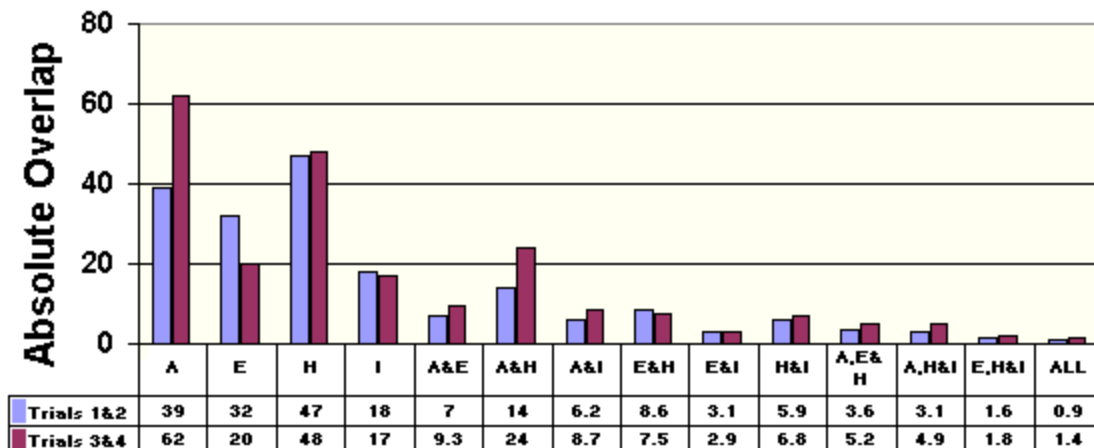


Fig. 2. Normalized estimates for all intersections (expressed as a percentage of total joint coverage).

We can see from Fig. 2 that in November 1997, only 1.4% of all URLs indexed by the search engines were estimated to be common to all of them. The maximum coverage by any search engine was by AltaVista, which had indexed an estimated 62% of the combined set of URLs. In July 1997, HotBot was the largest, with 47% of the combined set of URLs, and the four engine intersection was only 0.9%.

The ratios we obtained seem consistent with other estimates of the sizes of these search engines at the time. [Search Engine Watch](#) reported the following search engine sizes (as of November 5, 1997): AltaVista = 100 million pages, HotBot = 80 million, Excite = 55 million, and Infoseek = 30 million

pages. The October 14, 1997 [press release](#) from AltaVista also states a 100 million page database. Hence if the size of AltaVista was 100 million pages, then the total coverage of all the search engines should have been about $100 \text{ million} / 0.62$, or roughly 160 million pages in November 1997 (see [Fig. 3](#)). Similarly, the estimate that the four-engine intersection was 1.4% leads us to estimate that only roughly 2.25 million pages were common to all search engines in November 1997.

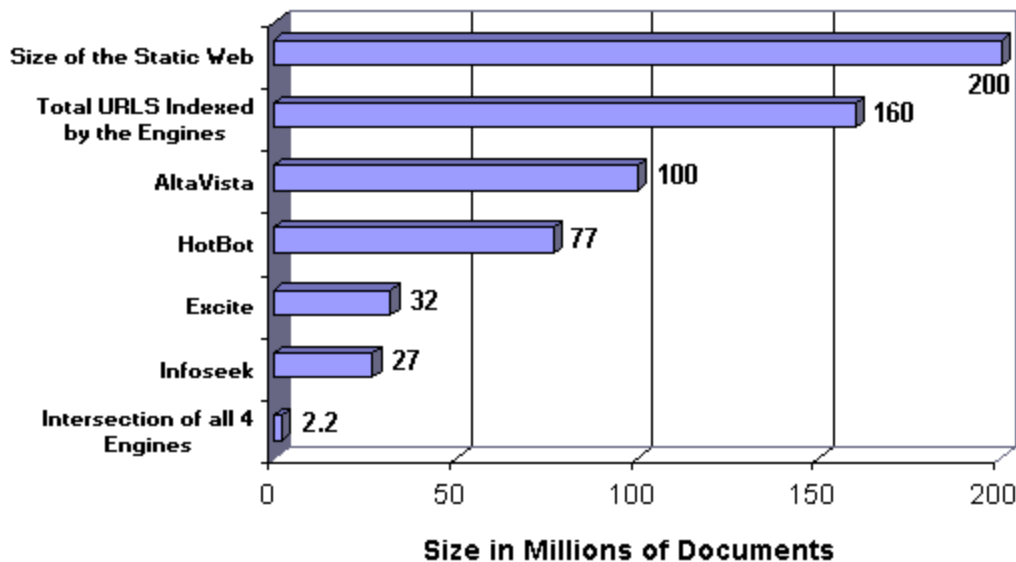


Fig. 3. Absolute size estimates for November 1997.

Further, we observe that each search engine seems to have indexed roughly a constant fraction of every other search engine. For example, from trials 3 and 4, it would seem that AltaVista indexed 50% of every search engine, and Infoseek indexed 15%, and so forth. This seems to suggest that the indices were constructed independently, and that these fractions are in general representative of each engine's coverage of the Web at large. On this assumption, if AltaVista had indexed 50% of the Web in November 1997, then the static portion of the Web could be estimated to be at least 200 million pages. Note that all engines have a bias towards well-connected, content-rich pages. The probability that AltaVista indexed an arbitrary page on the Web is likely to have been less than 0.5. Hence, the true size of the static Web was probably larger than 200 million documents in November 1997. This is in the same range as the estimate by Z. Smith, which was based on a crawl done by [Alexa](#) that found 80 million public pages in January 1997 and a predicted annual doubling rate [4].

6. Conclusion

In this paper we describe the first attempt to measure the coverage and overlap of public Web search engines using a statistically sound sampling technique. Our estimation scheme works on the basis of individual URL containment and thus we can estimate intersections between engines as well. Our measurements taken in November 1997 provided size ratios that were consistent with third party measurements reported in the press. We found AltaVista to be the largest search engine at that point in time with a 62% share of the combined set of URLs indexed by the four major engines, and a consistent 50% coverage of each of the three other search engines. Based on estimates that AltaVista's size was approximately 100 million documents, we conjecture that the size of the static public Web as of November 1997 was at least 200 million documents.

Our approach for search engine comparison has a clear statistical objective basis. Although it gives a higher weight to long, content rich pages in the language of the the lexicon, this bias is well understood and it is in principle computable for every page on the Web. Thus we have a method that can be used by any interested party to estimate the coverage of publicly available engines. Further, by modifying the lexicon suitably the method can be biased towards estimating coverage of a particular language or even a (broad) topic.

We are exploring the possibility of transferring this technology to a third party interested in providing a periodic evaluation of search engine coverage.

References

- [1] D. Brake, Lost in Cyberspace, *New Scientist*, June 28, 1997, <http://www.newscientist.com/keysites/networld/lost.html>
- [2] D. Brake, A. Broder, S. Glassman, M. Manasse, and G. Zweig, Syntactic clustering of the Web, in: *Proc. of the 6th International World Wide Web Conference*, April 1997, pp. 391–404, <http://www6.nttlabs.com/papers/PAPER205/PAPER205.html>.
- [3] A. Sinclair, *Algorithms for Random Generation and Counting: A Markov Chain Approach*. Birkhauser, 1993.
- [4] Z. Smith, The truth about the Web; crawling towards eternity, *Web Techniques Magazine*, May 1997, <http://www.webtechniques.com/features/1997/05/burner/burner.shtml>

URLs

- [Literature About Search Services](http://www.ub2.lu.se/desire/radar/lit-about-search-services.html) <http://www.ub2.lu.se/desire/radar/lit-about-search-services.html>
- [Alexa](http://www.alexa.com/) <http://www.alexa.com/>
- [AltaVista](http://altavista.digital.com/) <http://altavista.digital.com/>
- [HotBot](http://www.hotbot.com/) <http://www.hotbot.com/>
- [MetaCrawler](http://www.metacrawler.com/) <http://www.metacrawler.com/>
- [Yahoo!](http://www.yahoo.com/) <http://www.yahoo.com/>
- [Search Engine Watch](http://searchenginewatch.com/size.htm) <http://searchenginewatch.com/size.htm>
- [Excite Size Estimate](http://www.excite.com/ice/size.html) <http://www.excite.com/ice/size.html>
- [Melee's Indexing Coverage Analysis](http://www.melee.com/mica/index.html) <http://www.melee.com/mica/index.html>
- [AltaVista press release](http://www.altavista.digital.com/av/content/pr101497.htm) <http://www.altavista.digital.com/av/content/pr101497.htm>

Vitae



[Krishna Bharat](#) is a member of the research staff at Digital Equipment Corporation's Systems Research Center. His current research interests include Web content discovery and retrieval, user interface issues in automating tasks on the Web, and speech interaction with hand-held devices. He received his Ph.D. in Computer Science from Georgia Institute of Technology in 1996, where he worked on tool and infrastructure support for building distributed user interface applications.



[Andrei Broder](#) has a B.Sc. from Technion, Israel Institute of Technology and a

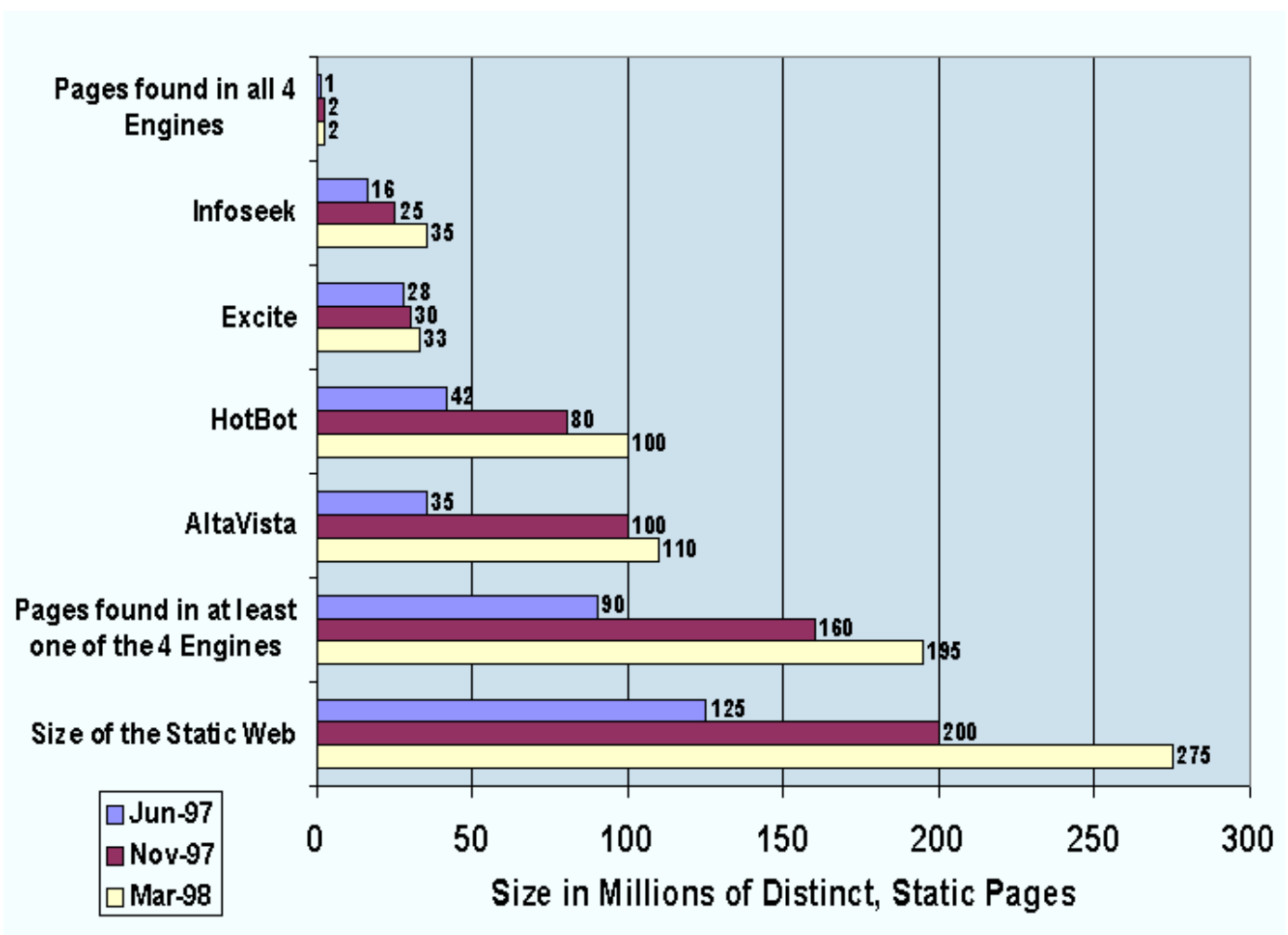
M.Sc. and Ph.D. in Computer Science from Stanford University. He is a member of the research staff at Digital Equipment Corporation's Systems Research Center in Palo Alto, California. His main interests are the design, analysis, and implementation of probabilistic algorithms and supporting data structures, in particular in the context of Web-scale applications.



Measuring the Web

[Krishna Bharat](#) and [Andrei Broder](#)
[DIGITAL](#), [Systems Research Center](#)
 {bharat, broder}@pa.dec.com

This is an update to our paper: [Estimating the Relative Size and Overlap of Public Web Search Engines](#) (©Elsevier Science), which was presented at the [7th International World Wide Web Conference \(WWW7\)](#) in April '98.



In March '98 we conducted a new experiment to measure four popular search engines. We employed 20,000 random queries using a lexicon of words drawn from a wide range of Web topics. The queries were used to sample random pages from the index of each search engine. Each page was then checked for containment within other engines. From the overlap estimates we derived estimates for the coverage of each engine, and based

on the known size of AltaVista, a lower bound for the size of the Web.

Our estimate is that as of March 1998 there were at least 275 million distinct², static³ pages on the web. (This is based on the fact that AltaVista at the time contained about 110 million distinct pages.)

Our data indicates that at that time AltaVista had the largest coverage (110 million distinct pages), having indexed 40% of the distinct pages on the Web. HotBot was second with a 36% coverage (100 million distinct pages). Infoseek and Excite covered 12% each (33 million distinct pages).

According to our experiments, the size of the static web was about 125 million in mid-97. It grew to about 200 million distinct pages by November 97 and to about 275 million distinct pages as of March 98. This means that the Web doubled in size in less than 9 months and is currently growing about 20 million pages per month.

Many people have asked how our study compares with the work of S. Lawrence and C. L. Giles from [NEC Research Institute](#) that was recently published in [Science](#). Their findings are broadly consistent with ours. However, we think that they overestimated the size of the Web in December. One reason might be that we try to count only pages that are distinct content-wise, while they count distinct normalized URLs. So, for instance if the same page appears in two locations it will be counted only once in our estimates but may be counted twice in theirs.

We believe our technique to be more accurate as well. We sampled 20,000 random pages per measurement which gives us a better statistical sample (less correlation) than their 575 sets of result pages. (Their size-of-the-web estimate is based on 302 sets of results.) We used an automatic technique to frame random queries based on a lexicon of 400,000 words covering a wide range of Web topics and languages. The lexicon was build from the vocabulary of 300,000 pages present in the [Yahoo!](#) hierarchy. In contrast, the NEC study employed queries used within NEC. We avoided using queries from query logs, because that tends to bias the sampling towards the interests of a particular group of users. Random queries give a more uniform sampling. Hence, we feel that our survey is more general and reflects better the true state of the World Wide Web at the time of the study.

-
1. The chart represents estimated coverage by various search engines and combinations thereof for distinct, static pages. Three sets of measurements were done — in June 1997, November 1997, and March 1998. Each measurement involved 20,000 random URLs, except the June 1997 measurement which involved 15,000.
 2. Pages are considered *distinct* only if they differ in content.
 3. *Static* pages are pages with content that is relatively static and can be reliably queried — as opposed to *dynamic*

pages, whose content changes continuously.

4. The absolute values are based on the information that AltaVista had indexed 35 million distinct pages in June 1997, 100 million distinct pages in November 1997, and 110 million distinct pages in March 1998. See this [paper](#) to understand the principles behind filtering duplicates in AltaVista.
 5. The size of the Web estimate is based on the observation that AltaVista had indexed roughly 28% of each of the other search engines in June 1997, roughly 50% in November 1997, and 40% in March 1998. We take these to be the probabilities of arbitrary pages on the Web being indexed by AltaVista. This is an underestimate because all search engines have a bias towards content-rich, well connected pages and towards pages on sites that engage in URL submission.
 6. For more details see: *Bharat, K. and Broder, A., [A technique for measuring the relative size and overlap of public Web search engines](#), in Proceedings of 7th International WWW Conference, Brisbane Australia.*
-

[Digital Systems Research Center](#)

130 Lytton Avenue, Palo Alto, CA
94301
Tel: (650) 853-2100 Fax: (650) 853-
2104



Send [comments](#) to the owner of this page.

Last modified: Friday, 24-Apr-98 10:38:31

PDT

[Legal notice](#)