

The Indexable Web is More than 11.5 Billion Pages

A. Gulli

Universita' di Pisa, Dipartimento di Informatica
gulli@di.unipi.it

A. Signorini

University of Iowa, Computer Science
alessio-signorini@uiowa.edu

ABSTRACT

In this short paper we estimate the size of the public indexable web at 11.5 billion pages. We also estimate the overlap and the index size of Google, MSN, Ask/Teoma and Yahoo!.

Categories and Subject Descriptors

H.3.3 [Information Storage And Retrieval]: Information Search and Retrieval

General Terms

Design, Experimentation, Measurement

Keywords

Search Engines, Index Sizes, Size of the Web

1. INTRODUCTION

What is the current size of the Web? At the time of this writing, Google claims to index more than 8 billion pages, MSN claims about 5 billion pages, Yahoo! at least 4 billion and Ask/Teoma more than 2 billion. Two sources for tracking the growth of the Web are [8, 7], although they are not kept up to date. Estimating the size of the whole Web is quite difficult, due to its dynamic nature. Nevertheless, it is possible to assess the size of the publically indexable Web. The indexable Web [4] is defined as "the part of the Web which is considered for indexing by the major engines". In 1997, Bharat and Broder [2] estimated the size of Web indexed by Hotbot, Altavista, Excite and Infoseek (the largest search engines at that time) at 200 million pages. They also pointed out that the estimated intersection of the indexes was less than 1.4%, or about 2.2 million pages. Furthermore, in 1998, Lawrence and Giles [6, 3] gave a lower bound 800 million pages. These estimates are now obsolete.

In this short paper, we revise and update the estimated size of the indexable Web to at least 11.5 billion pages as of the end of January 2005. We also estimate the relative size and overlap of the largest Web search engines. Precisely Google is the largest engine, followed by Yahoo!, by Ask/Teoma, and by MSN. We adopted the methodology proposed in 1997 by Bharat and Broder [2], but extended the number of queries used for testing from 35,000 in English, to more than 438,141 in 75 different languages. We remark that an estimate of the size of the web is useful in many situations, such as when compressing, ranking, spidering, indexing and mining the Web.

Copyright is held by the author/owner.
WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

2. ESTIMATING SIZES AND OVERLAPS

We review [2] and point out where our approach differs. The idea is quite simple: suppose we have two search engines A and B with size $s(A)$ and $s(B)$, respectively, and intersection $A \& B$. Let $Pr(A)$ represent the probability that an element belongs to the set A , and let $Pr(A \& B|A)$ represent the conditional probability that an element belongs to both sets given that it belongs to A . Then, $Pr(A \& B|A) \approx s(A \& B)/s(A)$ and similarly, $Pr(A \& B|B) \approx s(A \& B)/s(B)$, and therefore the relative size is $s(A)/s(B)$ that is approximately $Pr(A \& B|B)/Pr(A \& B|A)$. The methodology estimates $\frac{s(A)}{s(B)}$ by the ratio between the fraction of URLs sampled from B found in A and the fraction of URLs sampled from A found in B . It also estimates the overlap (fraction of search engine A index, indexed by search engine B) as fraction of URLs sampled from A found in B .

To implement this idea, one needs a procedure for picking pages uniformly at random from the index of a particular engine - i.e. a *sampling procedure* -, and a procedure for determining whether a particular page is indexed by a particular engine - i.e. a *checking procedure*. We refer to [2] for a discussion on the bias of this methodology.

Sampling: Bharat and Broder suggested a query-based sampling procedure. A set of disjunctive and conjunctive queries is submitted to the target search engine and an URL at random is selected from the top 100 results. They built the query lexicon by indexing 300,000 documents, extracted from the Yahoo! directory, and creating a lexicon of about 400,000 terms. The terms were combined to form 4 trials for a total of 35,000 queries.

We extend this approach, by indexing the whole DMOZ.com directory - more than 4 million pages - and obtaining a set of 2,190,702 terms. DMOZ directory contains pages in more than 75 languages, unlike portion the Yahoo! directory used by Bharat & Broder. We sorted the terms by occurrence and divided the sorted list in blocks of 20 terms. From each block we extract one query term for each search engine in the test bed, obtaining a total of 438,141 one-term queries. Queries were divided in many trials and submitted to Google, MSN, Yahoo! and Ask/Teoma. For each query, at least one URL at random was selected from the first 100 results.

Checking: Bharat and Broder suggested a query-based checking procedure. For each URL u , they extracted the k most discriminant terms from the downloaded web page. Then, these k terms were submitted to each search engine to check if it can find u . They tested 35,000 URLs.

We adopted a simplified form of checking. In fact, every engine in our test bed provides an interface to check directly

if an URL is indexed. Unfortunately this requires having a well-formed URL to check: thus, a lot of care was taken to normalize URLs. In particular, we exploited several heuristics for checking dynamically-generated Web pages and we filtered those pages not recognized by the originating engine after normalization. As a result, the effective number of checked URLs was 486,752. Experimental results confirmed that this data set¹ is large enough to provide, for each search engine, a stable estimate of its searchable Web coverage (see Figure 1). This also confirms that the use of one-term queries is reasonable.

Both for sampling and for checking we exploit Helios, a flexible and open source meta-search engine described in a companion paper [1]. Each sampling query was submitted to Helios, which forwarded it to the search engines in parallel. A similar process was used for checking the URLs. Our experiments were conducted on a cluster of 43 Linux servers, requiring about 70Gb of bandwidth and more than 3600 machine-hours. We included Google, MSN, Yahoo! and Ask/Teoma as test beds for our experiments, since these engines claim to have the largest indexes of the Web. Results were retrieved, parsed and saved locally.

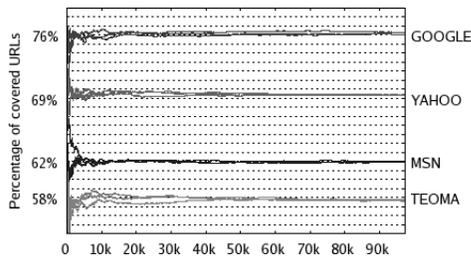


Figure 1: % URLs covered as the data set increases

Figure 1 shows that coverage does not change significantly as the number of checked URLs exceeds twenty thousand. Google covers around 76.2% of our sampling data set, Yahoo! covers around 69.3%, MSN covers around 61.9% and Ask/Teoma covers around 57.6%.

		T1	Rel.	T2	Rel.	T3	Rel.
Google	M	55.80%	1.41	55.27%	1.42	55.23%	1.42
	T	35.56%	1.65	35.89%	1.62	35.96%	1.62
	Y	55.63%	1.22	56.60%	1.20	56.04%	1.22
MSN	G	78.40%	0.71	78.48%	0.70	78.42%	0.70
	T	49.56%	0.87	49.57%	0.87	49.87%	0.86
	Y	67.38%	0.73	67.28%	0.73	67.30%	0.74
Ask/Teoma	G	58.83%	0.60	58.17%	0.62	58.20%	0.62
	M	42.99%	1.15	42.95%	1.15	42.68%	1.17
	Y	54.13%	0.84	53.70%	0.84	54.13%	0.83
Yahoo!	G	67.96%	0.82	67.71%	0.84	68.45%	0.82
	M	49.33%	1.37	49.38%	1.36	49.56%	1.36
	T	45.21%	1.20	45.32%	1.19	44.98%	1.20

Figure 2: Pairwise overlap, relative size (3 trials).

To reconcile the different pairwise size ratios, we used the least squares method to compute a best-fit solution on an overconstrained system. In particular, we estimated the engine sizes so that the sum of squared differences between the resulting estimates for the pairwise overlap was minimized.

Another possible approach to estimate the engine size uses linear programming. The objective function is to minimize the sum of the differences between the claimed engine sizes and size calculated using relative values. This results in 12 constraints like $size(a)/rel_size(a,b) - size(b) \leq d_i$ for

¹Due to space constraints, we report here the results of just three trials, for a total of 292,056 URLs. A more extensive description of these experiments, together with the data files, is available online [5]. The extended trials confirm the results given in this short paper.

$1 \leq i \leq 12$, where each d_i represent a pairing of engines. We use the declared search engine size as a lower bound for each engine variable. Our experiments showed that using just two lower bounds, the engine sizes are stable, confirming the results of the above least squares method.

Then, we expressed the engine size as a ratio respect the largest engine size (here, Google).

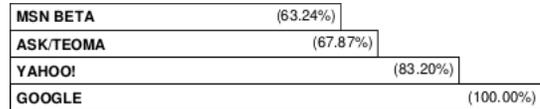


Figure 3: Estimated relative size per search engine

Figure 4 graphically represents the percentage of the indexable web that lies in each search engine's index and in their respective intersections.

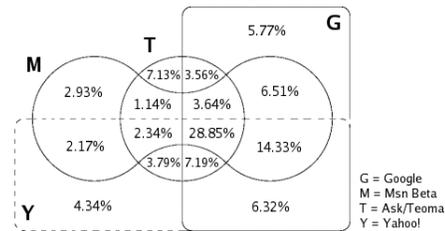


Figure 4: Results distribution across engines.

3. ESTIMATING INDEXABLE WEB

As suggested by [7], we assumed that the indexes of search engines were constructed independently. For any given engine E , we averaged the fraction of every other search engine index which appears also to be indexed by E . For instance, Google appears to index around the 68.2% of any other search engine, MSN index around 49.2%, Ask/Teoma index around 43.5% and Yahoo! index about 59.1%. We can consider these as representative of each engine's coverage of the indexable Web at large.

Furthermore, we can compute the size of the indexable Web by exploiting both the relative size estimated in section 2 and the absolute size declared by each search engine in our testbed. Averaging these values, we estimate the Indexable Web to be approximately 11.5 billion pages. As reported in Figure 4, the estimated intersection of all four indexes is 28.85%, or about 2.7 billion pages, and their union is about 9.36 billion pages.

Acknowledgement

We thank the Department of Computer Science at University of Iowa for providing access to computing resources and B. Codonotti, G. Del Corso, P. Ferragina, R. Gini, M. Parton, F. Romani and A. M. Segre for their helpful comments.

4. REFERENCES

- [1] A.Gulli and A. Signorini. Building an open source meta search engine. In *14th WWW*, 2005.
- [2] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. In *7th WWW*, 1998.
- [3] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999.
- [4] E. Selberg. *Towards Comprehensive Web Search*. PhD thesis, University of Washington, 1999.
- [5] <http://www.cs.uiowa.edu/~assignori/web-size/>
- [6] <http://www.neci.nj.nec.com/homepages/lawrence/>
- [7] <http://searchengineshowdown.com/stats/>
- [8] <http://searchenginewatch.com/article.php/2156481>