Volume 2, No. 3, May-June 2011
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info
ISSN No. 0976-5697

Indexing Size Approximation of WWW Repository with Leading Information Retrieval and Web Filtering Robots

Ijaz Ali Shoukat*, Mohsin Iftikhar , Abdul Haseeb
Department of Computer Science, College of Computer & Information Sciences,
King Saud University,
Riyadh, KSA.
ishoukat@ksu.edu.sa, miftikhar@ksu.edu.sa, akhan@ksu.edu.sa

***Abstract*:** The biggest information system of World Wide Web indexing is critical to estimate. Web is the beneficial and growing scientific utility like digital library to explore electronic literature to its lovers. Indexing estimation of WWW information is an open problem since 1998. Yahoo has claimed 19 billion web documents as its indexed size on which Google is not satisfied because in accordance with last published study by Gulli and Signorini the *total* "indexed web size" was around 11.5 billion pages. Web is growing hastily; what is the current size of web? Which search engine possesses large indexing of authentic information (PDF files)? Which search engine provides large indexing of all types of Web pages? This article provides the answers of all above questions. We estimated the index size of leading search engines (Google, Yahoo and MSN) under easy and cost effective approach because if easy way persists then why we select tough heuristics. Our technique relies on querying over the search engines with selected common affixes that can be a part of each and every document or web page. This paper concludes the total size of current "*indexed web contents"* and provides comparative analysis to support the scholars; which search engine has more authentic information and large indexing size.

Keywords: Index Size of Search Engines, Total Web Size, Comparison of Google, Yahoo and MSN, Web Crawlers, Web Robots

I. INTRODUCTION

Web estimation links with two categories; one is relative size of deep web data sources and other is actual index size estimation [1]. Web size estimation goes back to 1980 in which many heuristics are put forward like, measurements of total sites over the particular data sources, relative coverage techniques and query based sampling techniques. The query based sampling method is the actively adopted trend from 2003 to date [1]. To crawl and index whole web is critical issue. Search engines provide their best effort to cover over all web contents. New literature coverage and indexing procedure relies on two important concepts; (1) critical coverage that means every new information should be indexed and not be missed by search engines. (2) Critical availability that means new information should be available immediately after publishing [2]. The available searching crawlers can cover only 65% to 75% new published contents and take 5 to 13 days for getting index such type of information [2]. This paper estimates the indexed web pages and PDF documents over the three leading search engines (Google, Yahoo, MSN) to support the end users and scholars, which document repository is rich in web contents.

II. LITERATURE REVIEW

Web size estimation history goes back to 30 years old as it had been started in 1980. Here, we have only discussed the papers of last years that are mostly related to our work. In 2007, a set of query is randomly selected and a web monitoring system (WebMon) was used to estimate the coverage of new information relies on search engines as shown in a Table 1 below [2]. According to this study, search engines still miss 23% ~ 35% newly published information where the Meta search engines can cover 86.4% new

information that means the coverage can be improve from 9.4% to 21.6% by utilizing the meta crawlers.

Table 1 Newly Published information Coverage [2]

Site Types	Google	Yahoo	Msn	Collective Results
Top Five Sites	65.6%	61.8%	41.4%	77.0%
others	81.3%	66.0%	75.3%	89.9%

According to study [3] Northern Light possesses largest indexing data base with only 16% coverage and Altavista covers 50% of web content's indexing where the web contents are estimated 800 million in July 1999. Furthermore this study reported the following information as discussed in Table 2 below.

Table 2 No. of Users and Pages Visited[3]

Search Engines	Registered Users	Average Pages Visited per days
Yahoo	more than 100 million	465 million
Excite	51 million	123 million

In 2004 a study [4] is conducted to find out which search engine provides strong information coverage by utilizing statistical methods to check overlap in search engines by

issuing random queries. This study estimated the information coverage of the following search engines as discussed in a Table 3 below that are based on the data which is taken in November 1997 by some other sources.

Table 3 Information Coverage (Nov. 1997) [4]

Search Engines	Information Coverage (%) on November 1997
Hotbot	77 million
Altavista	100 million
Excite	32 million
Infoseek	17 million
Total Coverage	200 million pages

Finally, this study reported that the size of web is roughly 200 million documents on November 1997 in which Altavista had greater share of 62 % web information.

Table 4 Shared Percentage of storage Servers [5].

Types of Servers	Stored Information Percentage
Scientific/Educational	6%
Government	1.1%
Health	2.8%
Personal	2.3%
Community	1.5%
Religion	0.8%
Societies	1.9%
Pornography	1.5%

According to Steve Lawrence and C. Lee Giles, the indexing capability of search engines vary and they may not be able to index new information even for months and they claims that, still no search engine is able to index more than 16% of the overall web [5] . furthermore, this study reported that world Wide Web has been increased to 800 million pages with 6 terabytes of text data and the no. of running web servers are about 3 million in which each type of server has different ratio of shared percentages over the web as discussed in a Table 4 . In 2005, Gulli and Signorini adopted random samples of URLs to measure the overlap between different search engines by calculating the fractional ratio of URLs between the two selected search engines. This study [6] reported that the total size of indexed web is 11.5 billion pages approximately in which the share of Google, Yahoo, MSN and Ask is shown in a Table 5 below.

Table 5 Index Size in 2005 [6]

Search Engines	Indexing Size (Billions) in 2005
Google	8 Billion
Yahoo	4 Billion
MSN	5 Billion
Ask	More than 2 Billion

Many studies [7], [8], [9], [10] related to web content's estimation have been reported; Lawrence and Giles in 1998

use a method to find out the relative size of data source based on overlapping investigation of two selected samples. In 2003 the Bolshakov and Galiciahar use query analysis approach to estimate the total no. of pages written in specific language that are possessed by Google.

III. METHODOLOGY

Web can deeply be analyzed only by utilizing query interface (Bergman 2001). We estimated the web population with two point of views: (1) indexing size of all types of web pages to find out the size of current web indexed pages. (2) Indexing judgment of PDF documents resides on the web and indexed by search engine(s) to investigate which search engine possesses large repository of quality oriented documents. We followed the following estimation approach.

A. Estimation Approach

Our motive is to estimate the size of indexed web under easy and cost effective approach which can easily be implemented by end user side. Many people used benchmarking tools for this purpose which actually rely on query implementation approach. We did not use any bench marking tool because if easy way persists then why we adopted the complex one. Our motive is to get actual results in easy way from end user side but not to practice different kinds of third party tools. Our method follows universal affixes like article ("and"), indefinite article ("a"), definite article ("the"), verbs ("is", "has", "have"), propositions ("in", "of", "to", "for") and other common keywords like ("Abstract", "Keywords", "introduction", "results", "conclusion", "study", "paper", "table") which cannot be neglected while writing a document, web page or paragraph even a line. We selected these all keywords and queried them over Google, Yahoo and MSN by using their advance search functions to get total no. of results as shown against each keyword one by one as summarized in Table 7. Similarly, for analyzing the probability relationship of PDF documents indexed by Google, Yahoo, and MSN, we queried the same selected keywords one by one by selecting the advance search option "filetype:pdf". The summarized statistical relationship of PDF document's probability among the search engines is discussed in Table 9.

IV. PRACTICAL RESULTS

The approximate number of results of all type of documents and web pages against each type of keyword are reported in Table 6, 7 and Table 8 respectively for Google, Yahoo and MSN. The results are taken for the period of 6 months starting from October 2010 to March 2011. During this period the results contain minor variations due to newly indexed and deleted documents. For newly indexed and dead link deleted documents; we used variation factor $\alpha(G)$, $\beta(Y)$ and $\omega(M)$ for Google, Yahoo and MSN respectively. The minor updating variations in future indexing have not immediate effect in left most large numeric value because in this study our motive is to find out approximated *indexed web* contents for the year of Jan, 2011 that's why we have neglected the future incremental/ decremented factors $\alpha(G)$, $\beta(Y)$ and $\omega(M)$ for all selected search engines

Table 6. Google's Approximated Index Estimation Results for ALL File Types

Common Words	Oct – Nov 2010	Dec – Jan 2011	Feb – March 2011
"a"	25,300,000,000 ± α(G)	25,310,000,000 ± α(G)	25,270,000,000 ± α(G)
"the"	25,310,000,000 ± α(G)	25,330,000,000 ±α(G)	25,270,000,000 ± α(G)
"in"	25,300,000,000 ± α(G)	25,310,000,000 ± α(G)	25,270,000,000 ± α(G)
"and"	25,310,000,000 ± α(G)	25,330,000,000 ±α(G)	25,270,000,000 ± α(G)
"of"	25,305,000,000 ± α(G)	25,320,000,000 ± α(G)	25,270,000,000 ± α(G)
"to"	25,300,000,000 ±	25,310,000,000 ± α(G)	25,270,000,000 ± α(G)
"for"	25,270,000,000 ± α(G)	25,270,000,000 ± α(G)	25,270,000,000 ± α(G)
"this"	25,290,000,000 ± α(G)	25,290,000,000 ± α(G)	25,270,000,000 ± α(G)
"is"	25,300,000,000±α(G)	25,310,000,000 ± α(G)	25,270,000,000 ± α(G)
"has"	6,410,000,000 ± α(G)	6,410,000,000 ± α(G)	9,270,000,000 ± α(G)
"have"	7,180,000,000 ± α(G)	7,180,000,000 ± α(G)	10,810,000,000 ± α(G)
"abstract"	168,000,000 ± α(G)	198,000,000 ± α(G)	387,000,000 ± α(G)
"key words" OR "keywords"	423,000,000 ± α(G)	453,000,000 ± α(G)	2,290,000,000 ± α(G)
"introduction"	375,000,000 ±α(G)	4275,000,000 ± α(G)	489,000,000 ± α(G)
"conclusion"	118,000,000 ± α(G)	143,000,000 ± α(G)	177,000,000 ± α(G)
"results"	2,350,000,000 ± α(G)	3,347,000,000 ± α(G)	4,360,000,000 ±α(G)
"study"	851,000,000 ±α(G)	991,000,000 ± α(G)	1,120,000,000 ± α(G)
"paper"	1,030,000,000 ± α(G)	1,330,000,000	1,950,000,000 ± α(G)
"table"	969,000,000 ± α(G)	371,000,000 ± α(G)	395,000,000 ± α(G)
Selected large Value	**More than 25 ± α(G) Billions**		

Table 7 Yahoo's Approximated Index Estimation Results for ALL File Types

Common Words	Oct – Nov 2010	Dec – Jan 2011	Feb – March 2011
"a"	7,000,000,000 ± β(Y)	7,600,000,000 ± β(Y)	7,100,000,000 ± β(Y)
"the"	7,010,000,000 ± β(Y)	7,600,100,000 ± β(Y)	7,170,000,000 ± β(Y)
"to"	7,580,000,000 ± β(Y)	7,590,000,000 ± β(Y)	7,070,000,000 ± β(Y)
"in"	7,230,000,00 ± β(Y)	7,030,000,00 ± β(Y)	6,970,000,000 ± β(Y)
"and"	7,310,000,000 ± β(Y)	7,010,000,000 ± β(Y)	6,810,000,000 ± β(Y)
"of"	7,480,000,000 ± β(Y)	6,880,000,000 ± β(Y)	6,630,000,000 ± β(Y)
"for"	6,980,000,000 ± β(Y)	6,370,000,000 ± β(Y	6,240,000,000 ± β(Y)
"this"	5,000,000,000 ±β(Y)	5,001,000,000 ± β(Y)	4,680,000,000 ± β(Y)
"is"	5,470,000,000 ± β(Y)	5,270,000,000 ± β(Y)	4,960,000,000 ± β(Y)
"has"	2,370,000,000±β(Y)	2,270,000,000 ± β(Y)	2,130,000,000 ± β(Y)
"have"	3,530,000,000 ±β(Y)	3,430,000,000 ± β(Y)	3,360,000,000 ± β(Y)
"abstract"	50,000,000±β(Y)	47,000,000 ± β(Y)	44,500,000 ± β(Y)
"key words" OR "keywords"	415,000,000 ± β(Y)	341,000,000 ± β(Y)	212,000,000 ± β(Y)
"introduction"	221,000,000 ± β(Y)	201,000,000 ±β(Y)	194,000,000 ± β(Y)
"conclusion"	44,400,000 ± β(Y)	40,400,000 ± β(Y)	41,100,000 ± β(Y)
"results"	658,000,000 ± β(Y)	656,000,000 ± β(Y)	618,000,000 ± β(Y)
"study"	256,000,000 ± β(Y)	251,000,000 ± β(Y)	235,000,000 ± β(Y)

"paper"	259,000,000 ± β(Y)	248,000,000 ± β(Y)	240,000,000 ± β(Y)
"table"	371,000,000 ± β(Y)	343,000,000 ± β(Y)	336,000,000 ± β(Y)
Selected large Redundant Value	= **More than 7 ± β(Y) Billions**		

Table 8 MSN's Approximate Index Estimation Results for ALL File Types

Common Words	**Oct – Nov 2010**	**Dec – Jan 2011**	**Feb – March 2011**
"a"	11,800,000,000 ± ω(M)	11,100,000,000 ± ω(M)	11,300,000,000 ± ω(M)
"to"	11,590,000,000 ± ω(M)	11,010,000,000 ± ω(M)	11,100,000,000 ± ω(M)
"in"	7,680,000,000 ±ω(M)	8,300,000,000 ± ω(M)	7,300,000,000 ± ω(M)
"the"	10,400,000,000 ± ω(M)	9,400,000,000 ± ω(M)	7,290,000,000 ± ω(M)
"and"	7,760,000,000 ± ω(M)	9,300,000,000 ±ω(M)	10,400,000,000 ± ω(M)
"of"	7,370,000,000 ± ω(M)	8,100,000,000 ± ω(M)	10,700,000,000 ± ω(M)
"for"	7,070,000,000 ± ω(M)	7,800,000,000 ± ω(M)	9,480,000,000 ± ω(M)
"this"	5,110,000,000 ± ω(M)	4,100,000,000 ± ω(M)	4,930,000,000 ±ω(M)
"is"	5,820,000,000 ± ω(M)	7,000,000,000 ± ω(M)	9,000,000,000 ± ω(M)
"has"	2,390,000,000 ±ω(M)	2,190,000,000 ± ω(M)	2,180,000,000 ± ω(M)
"have"	3,740,000,000 ±ω(M)	3,470,000,000 ± ω(M)	3,470,000,000 ± ω(M)
"abstract"	48,600,000 ± ω(M)	47,300,000 ± ω(M)	45,200,000 ± ω(M)
"key words" OR "keywords"	205,000,000 ±ω(M)	204,000,000 ± ω(M)	204,000,000 ± ω(M)
"introduction"	210,000,000 ± ω(M)	201,000,000 ± ω(M)	197,000,000 ±ω(M)
"conclusion"	47,800,000 ± ω(M)	40,800,000 ± ω(M)	39,500,000 ± ω(M)
"results"	706,000,000 ±ω(M)	702,000,000 ± ω(M)	640,000,000 ± ω(M)
"study"	269,000,000 ± ω(M)	250,000,000 ± ω(M)	246,000,000 ± ω(M)
"paper"	275,000,000 ±ω(M)	255,000,000 ±ω(M)	246,000,000 ± ω(M)
"table"	395,000,000 ± ω(M)	363,000,000 ± ω(M)	345,000,000 ±ω(M)
Selected large Redundant Value	= **More than 11 ± ω(M) Billions**		

To find out the quality oriented document's probability judgment, we practically implemented query method with selected keywords queried one by one over selected search engines. We have observed the following probability relationship between Google and Yahoo for PDF document's indexing judgment as summarized in Table 9 below.

Table 9 PDF Document's Probability

Search Engines	Document's Probability relationship
Google	$\mathbb{P}\{G\sum(ó)\} > \mathbb{P}\{Y\sum(ó)\}$
Yahoo	$\mathbb{P}\{Y\sum(ó)\} < \mathbb{P}\{G\sum(ó)\}$
MSN	NA. as it does not provide facility to refine search with PDF file filtering option
Abbreviations	G represents Google Y represents Yahoo $\mathbb{P}$ represents Probability

	∑(ó) represents total no. of PDF results

The no. of results against all types of files are shown in a Table 9 below.

Table 10 All Files Index Size

Search Engines	Index Size in Billions
Google	25 ± α(G) Billions
Yahoo	7 ± β(Y) Billions
MSN	11 ± ω(M) Billions

Where the α(G), β(Y) and ω(M) are the future variation in the search engine's indexing database against newly indexed / deleted document(s) which we neglected at that moment to get valid results for Jan, 2011.

V. CONCLUSION

Web content population is swiftly expanding day by day. Due to this exponential growth, the web indexing size estimation is critical to investigate both quality of information as well as all type of basic information resides over WWW repository.

We have judged PDF indexing probability and total indexed web pages by implementing easy and cost effective approach. According to our practical observation based statistical results of Table 9, Google possesses more authentic documents (research studies PDF or HTML) than Yahoo and MSN does not provide advanced search option to filter PDF extension files. On the other hand with respect to all type of web contents Google is superior to Yahoo and MSN as Google contains more than 25 billion index of web pages followed by MSN with 11 billion web pages index and Yahoo is on 3rd position with 7 billion index of web pages as shown in Fig. 1 and Table 10.

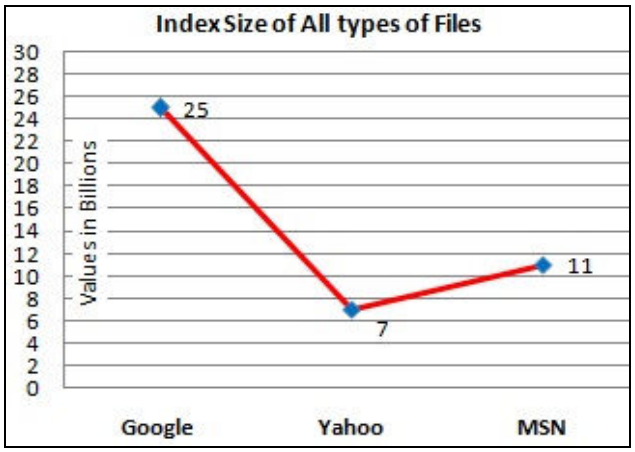

Fig. 1 Indexing Comparison of all types of Web Contents

The results of this article (Fig. 1) clearly invoke that the claim of yahoo ("*Yahoo has more than 19 billion indexed documents*") does not have any reality. Furthermore, from all three selected search engine's indexing comparison, the Google possesses largest indexing repository with more than 25 billion of web pages that represents; the size of *index able web* has been increased up to 25 billion of web pages by the January 2011.

VI. REFERENCES

[1] Liang. J. (2008), Estimation Methods for the Size of Deep Web Textural Data Source: A Survey, ACM Transactions on Computational Logic.

[2] Kim. Y. S. and Kang. H. B. (2007), Coverage and Timeliness Analysis of Search Engines with Webpage Monitoring Results,in WISE 2007, Springer LNCS 4831, Pages (361–372).

[3] Selberg. E. Etzioni. O. On the Instability of Web Search Engines, Go2Net, Inc. 999 Third Ave. Suite 4700, Seattle, WA 98104.

[4] Bharat. K. and Broder. A. (2004), A technique for measuring the relative size and overlap of public Web search engines, DIGITAL, Systems Research Center, 130 Lytton Avenue, Palo Alto, CA 94301, USA.

[5] Lawrence. S. and Giles. C. L. (1999), Accessibility of information on the web, NATURE, VOL 400, 8 JULY 1999. Macmillan Magazines Ltd.

[6] Gulli. A. and Signorini. A. (2005), The Indexable Web is More than 11.5 billion pages, *WWW 2005,* May (10–14), Chiba, Japan. ACM 1595930515/05/0005.

[7] Ahmad. R. , Ahmad. F. and Shah G. J. (2008), Overlap in Web Search Results: A Study of Five Search Engines, *Library Philosophy and Practice 2008.*

[8] Bar-Yossef. Z. and Gurevich. M. (2006), Random Sampling from a Search Engine's Index, *WWW 2006*, (May 23–26), Edinburgh, Scotland. ACM 1595933239/06/0005.

[9] LING, Y., MENG, X. AND LIU W. (2008). An attributes correlation based approach for estimating size of Web databases. *Journal of Software*, 19(2), 224-236.

[10] Lu. J. and Li. D. (2009), Estimating deep web data source size by capture–recapture method, Springer-Information Retrieval, Volume 13, Number 1, 70-95, DOI: 10.1007/s10791-009-9107-2009.