Multimodal Technologies and Interaction

MDPI

Article

Webometrics: Some Critical Issues of WWW Size Estimation Methods

Srinivasan Mohana Arunachalam *,†,‡, **Adamantios Koumpis** ‡ and **Siegfried Handschuh** ‡

Faculty of Informatics and Mathematics, University of Passau, 94032 Passau, Germany; adamantios.koumpis@gmail.com (A.K.); siegfried.handschuh@uni-passau.de (S.H.)

* Correspondence: mohana01@gw.uni-passau.de; Tel.: +49-151-6604-4823

† Current address: University of Passau, Innstrasse 43, 94032 Passau, Germany.

‡ These authors contributed equally to this work.

Received: 12 February 2018 ; Accepted: 30 March 2018 ; Published: 2 April 2018

Abstract: The number of webpages in the Internet has increased tremendously over the last two decades however only a part of it is indexed by various search engines. This small portion is the indexable web of the Internet and can be usually reachable from a Search Engine. Search engines play a big role in making the World Wide Web accessible to the end user, and how much of the World Wide Web is accessible on the size of the search engine's index. Researchers have proposed several ways to estimate this size of the indexable web using search engines with and without privileged access to the search engine's database. Our report provides a summary of methods used in the last two decades to estimate the size of the World Wide Web, as well as describe how this knowledge can be used in other aspects/tasks concerning the World Wide Web.

Keywords: search engines; index sizes; WWW size estimation; graph structure; webometrics

1. Introduction

The World Wide Web consists of millions of websites and billions of documents which are accessed through a search engine. The search engines can further be classified based on their coverage and who provides them. Search engines like Google, Yahoo!, Bing and Ask.com are at the top currently. Most people use Google as their primary search engine as shown in Figure 1. Google is mostly used because of it's coverage(android phones use Google as search engine by default) and because of the lack of awareness/popularity of the other search engines. Currently Google has the biggest index size, which means it covers a lot more of the World Wide Web than the rest of the search engines combined as appeared in [1] and in a more recent work in [2] which in turn cater to a wider audience. Webometrics (also Cybermetrics) is the study of quantitative aspects of the World Wide Web such as the number of hyperlinks and its Graph nature which in turn is used to study the social phenomena which characterize its evolution [3]. It also helps to study how the different search engines can have a competitive advantage over each other. One of the immediate reasons why Google dominates the other search engines is its index size, which is the number of documents it has indexed at a point in time. It is bigger than all the other search engines combined, which gives it a tremendous competitive advantage. What this means is that, Google covers a lot more of the Web than the search engines, attracting a wider audience. But only a big index size would not be very useful, as the web is filled with websites which are not valid are duplicates or contain very little useful information but still get indexed; on this see also [1,4,5].

In the following paragraphs we present how the estimation task of the WWW has helped in the field of Webometrics and then describe a list of methods that have been developed in the past to measure the size of the indexed World Wide Web, directly and indirectly.

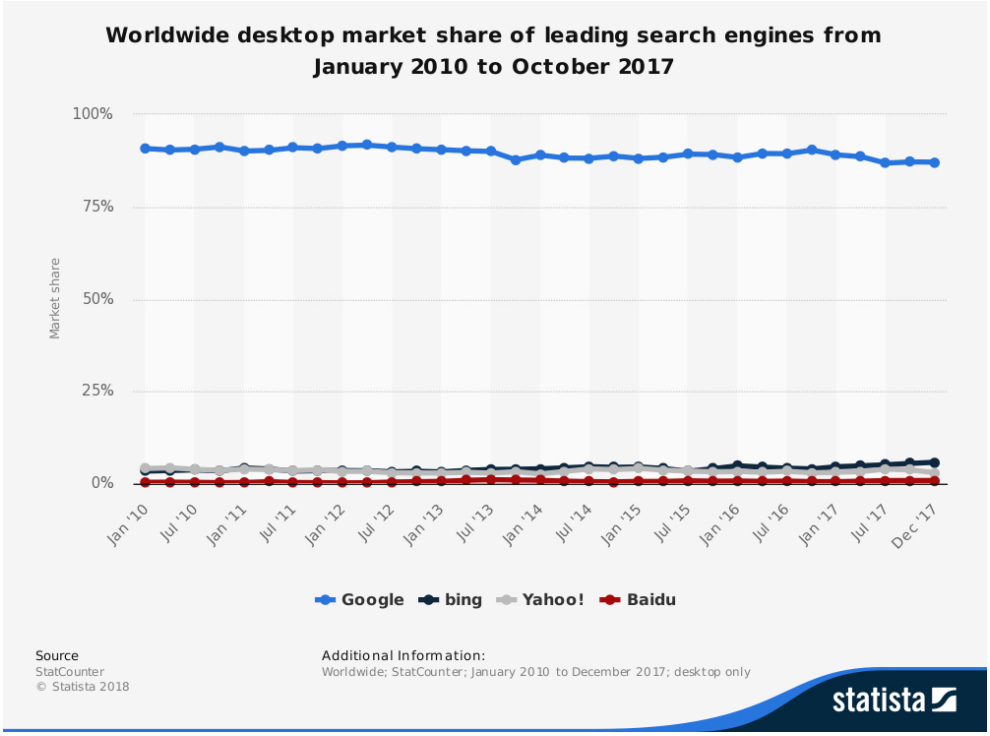

Figure 1. Search engine market share taken from statista.

2. Contributions to Webometrics Study

The following sections explain how the study of the World Wide Web has contributed to webometrics.

2.1. On Webometrics

Webometrics was one of the methodologies developed during mid 1990s concerning the quantitative aspects of information and met with confusion about its scope and relation to the other emerging fields namely Netometrics, Webometry, Internetometrics, Cybermetrics and Webometrics among others. This was due to the fact that all these methods were developed drawing on ideas from bibliometrics. Björneborn and Ingwersen in [6] redefined webometrics as a subfield of cybermetrics and sharing ideas with bibliometrics and scientometrics. They also extended on graph nature of the web and on Small-World in [7], which will be discussed in Section 2.4. Webometrics showed that the statistical analysis used in academic literature can also be applied to the study of link analysis in the World Wide Web and under controlled circumstances promising results can be obtained. This is in compliance also to (Vaughan et al., 2004) that aimed to measure coverage bias in search engines in terms of building a sample of sites and examining the possible causes in the differences of results appearing in search engines, and converging to the conclusion that it is the visibility of a site that affects its coverage by search engines, the latter being measured by the number of links to it.

2.2. Study of Overlap

As stated previously despite what common sense would dictate, even though all search engines attempt to cover all the content available on the Internet, their indexes reveal otherwise [8–10]. The majority of a search engine's index, around 85% is unique and shares very little (overlap of 3%) with the other search engines.

Bharat and Broder found that Altavista being the largest search engine at the time had an overlap of 50% with the other 3 search engines Excite, Hotbot and Infoseek. But this bigger overlap was since Altavista covered 62% of the distinct URLs in the crawl data [10].

The nature of overlap among search engine results was revisited in [8] using Google, Yahoo! and Ask Jeeves and considering only the links on the first page of results. They found that 85% of the links found in their tests were distinct, 12% of the results were shared by two of the three search engines and only 3% were shared by all three search engines under study. The resultant links were obtained by performing 10,316 random user entered queries taken from Dogpile.com and sampling the results from the first page of the results. This resulted in a total of 336,232 URLs of which they found 84.9% to be distinct [8].

The findings above reveal that not all search engine index the same content, and the websites which are shared by multiple search engines is very little if we go past the first results page. This presents an opportunity for meta search engines like Dogpile.com that retrieves from multiple search engines, specifically the top-ranking results from them, and presenting it to the user effectively reducing the time taken to find a piece of information on the WWW.

2.3. Graph Nature of the World Wide Web

The World Wide Web owing to its connected nature can be represented as a connected graph, where the websites are the vertices (or nodes) and hyperlinks are the arcs. The graph nature of the Web was studied by Broder in [10] and later in [3,11] with additional technical improvements. Broder and Kumar performed their experiments using the Connectivity Server 2 (CS2) at the Compaq Systems Research Center using crawl data from Altavista [11]. The graph is built by retrieving all the hyperlinks from a given page and then recursively retrieving all the hyperlinks from corresponding hyperlinks. The number of nodes in the graphs explodes within few levels in depth. This shows the highly connected nature of the websites. The Altavista crawl contains 203 million vertices (or websites) with 1433 million arcs. The study reveals the presence of a big strongly connected component (SCC) of about 56 million nodes and the rest of nodes forming two distinct groups with namely IN and OUT. The nodes in IN have edges leading to the nodes in SCC, and the nodes in OUT have edges coming from the nodes in SCC as shown in Figure 2. And the remaining nodes were grouped in TENDRILS which contain nodes which are connected only to the nodes in IN or OUT or completely disconnected. The traversals were obtained by using a BFS (Breadth-first) algorithm, SCC or WCC where the algorithm finds the Strongest Connected Component or the Weakest Connected Component respectively.

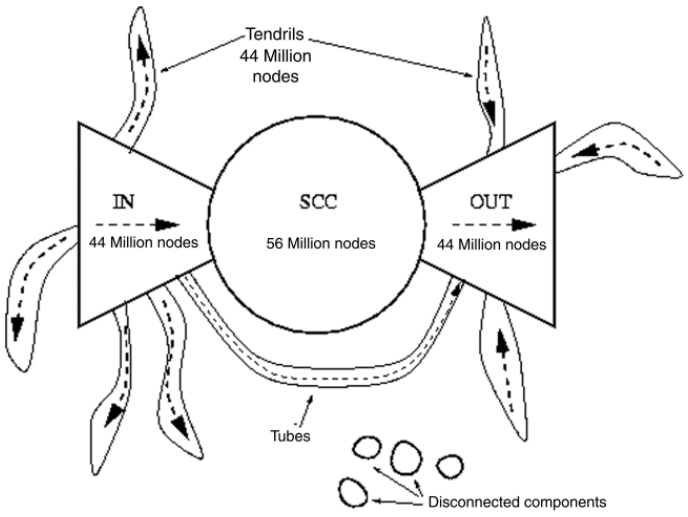

Figure 2. Connectivity of the web, as in [3].

2.4. Diameter of the Web Graph

Albert, Jeong and Barabasi first defined the distance between two randomly chosen nodes from the web graph as Diameter in [3,12]. The diameter is defined as the shortest distance between two

randomly chosen nodes in the web graph. The diameter is set to infinity in case a path doesn't exist. They found that on an average the diameter of the web is 19 and predicted that diameter could increase to only 21 over the next few years. They based their results on the analysis of the nd.edu domain, which consists of 325,729 nodes and 1,469,680 edges.

Broder and Kumar reexamined the above findings in [3] with the Altavista crawl and found evidence both supporting and critiquing the previous claim. In their graph analysis, they found that in a directed graph the average diameter was 16 and as an undirected graph, the average diameter was as low as 6. This was observed in the Altavista web graph in the form of SCC and IN/OUT relationship. This also proves the small world phenomena that there exists a short path between any randomly chosen node. They also found the probability that such a path exists is 24% for randomly chosen source and destination nodes.

2.5. Experiments on Closed Environment vs. the Web

We would like to mention similar experiments performed on specific parts of the Web before going into the method descriptions. In [12] we see experiments performed on all the links and documents in the the nd.edu domain to validate the diameter as proposed by Albert and Barabasi. And in recent years, work has been done to estimate the number of documents/articles in Google scholar in [13] and similarly to count academic articles from a number of sources in [14]. But what we want to point out here is that, these experiments are performed in a controlled environment, where a new document/article/entry into these databases are highly curated and studies by different metrics i.e., it is not difficult to find documents/articles in a library. In comparison, the WWW is highly diverse, and many similarities to the academic literature analysis may be superficial, as stated in [7].

3. Estimating the Size of the Indexed Web

The task of measuring the World Wide Web is to get information about the IP addresses which are currently in use and which are not. This can be done by trying to establish a connection with the IP address under consideration. A successful connection attempt would indicate the presence of a webservice or at least that it is unavailable (from a commercial perspective). An unsuccessful connection attempt would indicate a number of details about the webserver like e.g., if there is no webservice to consume the connection request, if the IP address is available for purchase (meaning its currently unused, which is very unlikely), if the server is not accepting connections (when a connection request comes from outside the local network), if an improper connection request, if the server is facing technical issues (this is hard to confirm during tests, as the server may be active at a later time without the tester's knowledge), etc.

Now that the task is defined, let's look at the test environment. We are trying to calculate the number of websites which are actively running on the World Wide Web. This means we must find all the websites which are accessed. Till 2008 the experiments only considered the IP addresses in the IPv4 address space, this however has expanded into the IPv6 scheme, since its first launch on 8 June 2011. This was done because of the growing size of the World Wide Web, the number of electronic devices and applications/services etc., demanding more and more IP addresses while the IPv4 address space ran out of publicly available IP addresses with the exception of unused IP addresses which are reserved by the Regional Internet Registries like IETF, AFRINIC and other organizations. The IPv4 address space contains 4 billion IP addresses, and so to find out the absolute size of the Web, we only need to probe the 4 billion IP addresses. This would have taken some few hundred years with the resources available back in 2002 [1]. Ignoring the impossible time constraints if we proceed to manually probe every single address, there are still other barriers which cannot be surpassed, which will be discussed in Section 2.2. And with the upgrade to IPv6, the previous input size is increased to 281 trillion (exactly: 281,474,976,710,767) IPv6 addresses. And so, most of the algorithms rely on statistics to estimate the size of the World Wide Web, which means that there is in fact no reliably "accurate" way anymore to measure the size of the Web and that we rely entirely on the use of statistical i.e., non-discrete methods.

3.1. Search Engines and WWW Size Estimation

A search engine's primary task is information retrieval but in the Internet a search engine has much higher responsibility as it is responsible for making accessible the content available on the Internet. A search engine does this by "crawling" the web to identify websites and index them [4,5,15]. Normally the crawling program called a "spider" does the crawling and indexing but a new website can also be directly introduced by manually submitting a website to a search engine's administrator/webmaster, who then determines the quality of the website and indexes it. This eliminates or shortens the time it takes for the crawler to manually identify the website and establish its value after which it gets indexed, or not. In this way we ensure that only websites which provide unique content get indexed; however, websites with redundant content frequently get indexed by abusing the method involved in the crawling process [5], and search engines regularly drop websites from their databases owing to a variety of reasons described as we describe below in Section 3.3. Such information is, in general, not available to the public; in fact, very little information about a search engine's index is available to the public and many researchers have developed multiple ways of using this very little publicly accessible information to estimate the size of the web. This paper provides an insight on some of those methods.

At any point in time, a search engine has a certain number of indexed pages in its index, this number varies over time. New websites are added, and websites are removed from index for reasons of ranking, relevancy, efficiency [2,16]. Different search engines have their own index management, so it is important that behavior of multiple search engines be observed. And each search engine has its own index, with its own crawler and or human powered directories [17]. Experiments across multiple search engines have showed a very low index overlap of 3% [8,9]. This means that over 85% of the results for a search query is unique for any given search engine [9,16]. This is quite contrary to what anyone would expect, i.e., if search engines try cover the content on the WWW, all of them would cover the same content. But the results have proved otherwise, as only a small percent of the indexable web is shared by multiple search engines, in a similar fashion as with floating icebergs that have a significant proportion of their mass below the surface of the water.

In table 1 we provide information about some web-based platforms giving information about the the web and multitude of analytical statistics about the web.

Table 1. Web-based platforms offering statistics.

Website	Information Provided
WorldWideWebSize	Daily estimates on Google and Bing index sizes
InternetLiveStats	Live update of variety of statistics on things connected to the internet
InternetWorldStats	Provides statistics on the world internet usage
Statista	Provides variety of statistics on the online and offline world of the consumer
Alexa	Provides commercial web traffic data and analytics
The Internet Map	Tries to display the web as a map
httpArchive	Provides statistics on how the data is constructed and served on the internet
Netcraft	Provides research data and analysis on many aspects of the internet

3.2. Methods Surveyed

In the following subsections we try do describe how in the past, researchers have tried to estimate the size of the web.

3.2.1. Statistical Approach Using Web Page Sampling

In 1998 Bharat and Broder [10] published their statistical approach to finding the coverage of a search engine relative to another search engine chosen. In their experiments, they considered Altavista, Hotbot, Excite and Infoseek search engines, with Altavista having the largest index size at 100 million pages reported by Search Engine Watch, Google was not yet "born" at that time. They performed the

experiments on two occasions on June & July 1997 and on November 1997. The experiment comprised of passing a randomly chosen query from a query set to all the search engines in consideration and retrieving the first 100 results [10]. The query set consists of disjunctive and conjunctive word pairs created from a lexicon of 400,000 words, which was extracted from 300,000 documents in the Yahoo! Hierarchy. The disjunctive (OR) queries of length 4, and the words chosen had roughly the same frequencies to reduce the bias towards words with high frequencies. The conjunctive (AND) queries of size 3 or less were used, as most search engines at that time provided 0 results for a conjunctive query of size 3. Currently we can have a conjunctive query up to 5 terms and we would still get hits provided the words are all in the same language and not imagined. A Google search would confirm this. Figure 3 Shows hit counts for a Google search with 10 terms combined with an AND. The terms used are Sentiment, Bacon, Alphabets, Memory, Chemicals, Lavender, Basket, Nature, Colosseum and Imagination.

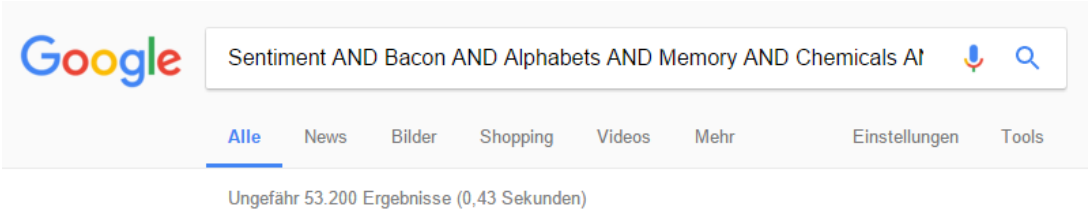

Figure 3. Hits counts for a conjunctive query of length 10.

The checking procedure involves finding out if a URL from one search engine's result set was found in another search engine's results. Multiple matching rules were used for checking the URLs, Full URL comparison, High similarity, Weak URL comparison. Four trials were conducted, trials 1 (10,000 queries) and 2 (5000) with disjunctive queries and trials 3 and 4 with conjunctive queries both of size 10,000. Sampling was done by randomly picking a URL from the top 100 results. The size estimates were obtained by calculating the percentage of overlap between two search engine results as conditional probability:

$$\frac{\text{Size(A)}}{\text{Size(B)}} = \frac{\Pr(\text{A \& B} \mid \text{B})}{\Pr(\text{A \& B} \mid \text{A})} \quad (1)$$

where Pr(A) and Pr(B) is the probability that an element belongs to set A and B respectively. Pr(A & B | A) and Pr(A & B | B) is the conditional probability that an element belongs to both sets given that it belongs to A or B respectively. The conditional probabilities are obtained by performing random sampling and checking of search engine results (URLs) as described above. Figure 4 shows the result of their experiment placing the number of static web pages in the World Wide Web at over 200 million pages.

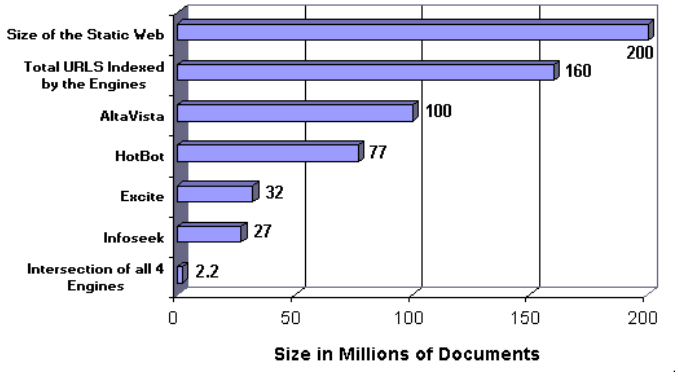

Figure 4. Size of static web pages in 1997 as in [10].

3.2.2. Updated Experiment Setting

In 2005 Gulli and Signorini revised and updated the findings of Bharat and Broder with an updated query lexicon built using the entire DMOZ directory [18]. Their lexicon consisted of 2,190,702 terms and ran the experiment for a total of 438,141 one term queries. Their experiments did not consider disjunctive or conjunctive queries. They estimated the size of the indexable web to be more than 11.5 billion which is the sum of the individual index sizes of the four search engines they considered Google, MSN, Ask/Teoma and Yahoo! after considering their overlap. 8 years since the first experiment, Altavista was no longer the most popular website and was subsequently purchased by Yahoo! in 2003, Yahoo! which was later acquired by Verizon in 2017.

3.2.3. Size Estimation through Quadrat Sampling

Population sampling is the process of taking a subset of subjects that is representative of the entire population. The sample must have sufficient size to warrant statistical analysis. Trudel and Rhodenizer applied the idea of population sampling to the Internet to determine the number of active web servers by manually probing every single IP address in each set of IP addresses to determining the number of active servers [1]. In the real world, a Quadrat would be a subsection of the total area under estimation, in comparison the Internet is the total area and a Quadrat would be a set of IP addresses, for example all the IP addresses in a given range (like a.b.c.0-255, has 256 IP addresses). The total number of Quadrats are more than sufficient to warrant a statistical analysis. Given all the IP addresses in a Quadrat they are manually probed to establish its availability. They developed a program to perform the probing and store the results and analyze them. Although every possible IP addresses in a Quadrat is a valid candidate for probing, a lot of the addresses are restricted or reserved for special purposes thereby making the extrapolation subject to inaccuracy. For example, there are reserved addresses like the addresses used for broadcast (255.255.255.255) or loopback (127.*.*.*), and then there are private networks (192.168.0-255.0-255) which are not used for routing over the Internet.

In their experiments, they probed 100 Quadrats through 100 runs, this covered a total of 25,600 IP addresses, of which 111 responded to a connection request, and 21,980 did not respond to a connection request and 3509 addresses were ruled out as being restricted and/or reserved. All requests were made on port 80 and a successful response, usually one that includes the HTTP success status message header 2XX, indicates the presence of a server. The average number of web servers per Quadrat was found to be 1.1 which multiplied over the total number of Quadrats (16,777,216 total Quadrats) gives an estimate of 18,622,710 active web servers in the IPv4 web address space. While the population estimation approach is interesting to implement in the IPv4 address space, it becomes near impossible in the newer IPv6 address space. This approach also does not factor in the presence of hosted servers, where a single physical machine can host multiple servers with the same address. Virtual hosting also introduces another problem, when multiple requests are made to the same machine hosting multiple virtual servers. This multiple probing could potentially be identified as an attack and all further probes could be blocked. Trudel and Rhodenizer implemented their idea in Perl to automate this process of probing, and displays the result of the scanning process [1].

3.2.4. Size Estimation through Extrapolation

Antal, Toine and Maurice [2] presented a novel approach to estimating the size of the indexable web using hit counts from search engine results. Their experiment was performed regularly over a period of 9 years through regularly performing a query search and retrieving the hit counts. The hit counts are the numerical counts displayed on the results page. Therefore, this method can be applied to any search engine provided it displays the hit counts. They've found that the hit counts are representative of the search engine's index, i.e., the number of documents a search engine has indexed at any point.

This was determined by carefully analysing the DMOZ corpus for word frequencies across the entire document collection. They could extrapolate the size of the document collection through the equation,

$$|C| = \frac{d_c(are) \times T}{d_T(are)} \tag{2}$$

where C is the size of corpus, d_c and d_T are the document counts for the word "are" in the two sample corpora C and T, in which T is the training corpus and C is the corpus to be measured.

The experiment involved passing a single word query to each of the search engines and retrieving the hit count using regular expressions from the respective results page. They targeted Google, Bing, Yahoo! and Ask.com but later stopped passing requests to Yahoo! and Ask.com owing to changes in the web page design which made retrieving the hit counts difficult. Also, Google and Bing provide custom APIs to their indexes which can be used in building a custom search engine which made making search requests and the hit counts retrieval easier. The query set consists of 28 terms taken from the DMOZ directory, they are high frequency pivot words and included foreign words [2]. The idea is that these high frequency words can be observed in all documents that are in the database. This experiment was scripted to automatically perform the search requests and retrieve the hit counts automatically for all the words and across the multiple search engines.

The experiment ran from March 2006 to January 2015 recording the estimated index sizes of Google and Bing as show in Figure 5 and the current results can be seen online at WorldWideWebSize.com. In Figure 5, we can see the variability in the index sizes of Google and Bing and each data point represents a average of 30 days, i.e., 15 days before and after a focus day as a window. The markers on top correspond to major updates made to their architectures of Google/Bing, obtained through public announcements posted by Google and Bing. The authors found correlation between the spikes/steps in the index and a corresponding update, for Google 20/24 updates correspond to spikes and steps and for Bing 6/12 matches. And for the cases where no match is found, they believe it is also due to architectural changes to the search engine, which was earlier suggested by Rousseau in his 1999 study.

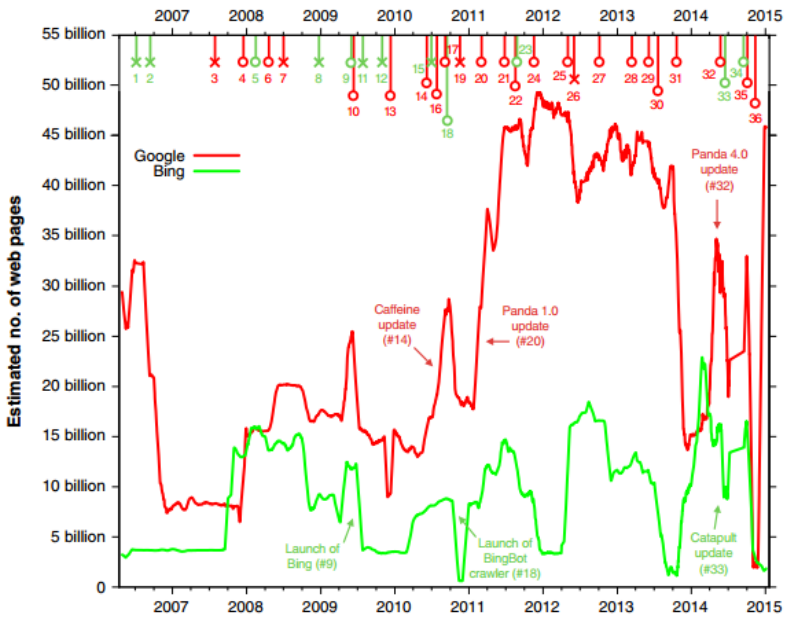

Figure 5. Estimated index sizes of Google and Bing as in [2].

3.3. Index Stability

As stated before, a search engine would frequently update its index, adding new valuable websites and dropping websites of less value. From Figure 5 it is visible that there are periods during when the index sizes have suddenly dropped and times when the index sizes have soared. Bar-Ilan [16] had previously performed a case study on the stability of a search engine's index where the search engines Altavista, Excite, Hotbot, Infoseek, Lycos and Northern Light were studied by random sampling hyperlinks from the first page of results to build a sample set representative of search engine's index. This experiment was performed 6 times subsequently, with the first 4 attempts to create a stable set of links and the next 2 attempts to test the stability (persistence) of the hyperlinks. During the last two attempts, it was found that Excite dropped and Lycos dropped more than 81% and 84% of their previously indexed hyperlinks respectively while Northern Light, a comparatively smaller search engine had dropped as little as 24% of its index [16]. In their study, an index was considered forgotten when a hyperlink which was previously retrieved in the first 4 attempts was not retrieved during the last 2 attempts. They also observed that some hyperlinks had disappeared during the second attempt but later reoccurred during the later attempts.

In [2] however stability of the entire index was studied through its size estimate over the 9 years. It was observed that the spikes in the index sizes were due to internal changes in the architecture of the index management as well as external changes such as the updates to the crawler algorithm (spider). But there are also points when the spikes could not be attributed to any event.

4. Discussion and Conclusions

In our paper we have presented methods used in estimating the size of the World Wide Web over the previous years and made a first attempt to assess their contribution to a study of what is called "Webometrics". The approaches described here do not require privileged access to a search engine's database and while the results are influenced by many biases, with sampling bias persistent across all the different methods. This sampling bias can be addressed by using the importance sampling over the Monte Carlo sampling technique [19]. While the statistical approach is useful to estimate the size of the indexable web, to get the absolute number of active websites on the internet, web crawling is the best way to do it. As we state in the paper, very little information about a search engine's index is available to the public.

Search engines themselves provide almost no information about their index while the statistics available are either result of work by network and telecommunication experts either with privileged access or simply access to network data. Further to this, we understand that at some extend this may be because such information is regarded as company secret, while it may be hard to know about the internal changes of the exact calculation methods applied to acquire the statistics unless one works for that company. The need for more transparency in this field may constitute one of the challenges of the World Wide Web Consortium as Web's main governance body.

Lewandowski [20,21] has studied in both papers the quality of web search engines. In particular, in the first study of 2006 the aim was to measure the frequency with which search engines update their indices and more specifically asking on the reason why one web page is updated within a day and some other within a week. This is a recurring theme also in the second study of 2008, where all search engines investigated exhibited shortcomings in updating their databases. Quite even worse, according to Lewandowski [21], 'none of the engines provides up-to-date copies even for the daily updated pages'. Apparently, this shortcoming affects the overall transparency in search engines as well as the trustworthiness of the entire area of Webometrics. In this context, the present study is neither offering the completeness of a review nor a new methodology towards estimating the size of the World Wide Web. And while we are aware that some of the presented approaches are now about 20 years old, and this, as expected, has implications for the current web architecture, it is still important to pinpoint the need for renewing the interest in this field.

Especially under the light of recent developments in the areas of Big Data, as well as the proliferation of the Web and its increasing value both at the commercial and at the social levels, the existence of independent bodies or third trusted parties that will have the responsibility, either in an enforced or in a self-governed approach, to take care of the aspects of 'measuring the Web' will be of paramount importance. Such a body would aim to get on a continuous base accurate and with high(er) certainty information and knowledge about the number and types of sites, pages and hyperlinks, helping continuously watch the evolving structure of the World Wide Web and its usage patterns.

It is obvious that counting the number of files or, alternatively, the number of URLs and web domains indexed is an indirect method. Further to this, it is known to us that there are proprietary file formats which are not covered by search engines. This might offer an opportunity to come up with new, disruptive and radically new methods that might use this as a challenge to improving the means and practices of the addressed field. However, it is easy to see that as we rely on estimates, and Google's indexing algorithm are not public, we are only able to speculate that the spikes in the index are due either to internal changes leading to new URLs being included or existing URLs to be dropped.

What we see as a need, and what we hope our article will contribute to is the revival of a discussion towards the establishment of an independent body—possibly under the World Wide Web Consortium (W3C) or the Internet Engineering Task Force (IETF) that will lead the effort to build a reliable and trustworthy observatory for what we called in the scope of the paper as 'Webometrics'. There are parallels that can be drawn between Webometrics and the field of 'traditional' census techniques. There, for example, there is a continuity since the ancient times that converged to approaches like the traditional census with yearly updates, the register-based census, the rolling census, etc. However, there have been innovations that have taken place like the use of post-enumeration surveys and what in census circles is called dual system enumeration (DSE). The latter is facilitated by computer matching techniques which can be automated, such as propensity score matching (PSM), namely a statistical matching technique that was first published by Rosenbaum and Rubin [22] and which attempts to estimate the effect of a treatment, policy, or any other intervention by accounting for the covariates that predict receiving the treatment. In our case of Webometrics, such covariates may be attributed, as mentioned above, either to new URLs being included or some pre-existing URLs to be dropped. Use of such techniques may not only increase the accuracy of the answers we can get, but may also help create a snowball effect in a variety of fields like cybersecurity and the protection of data infrastructures on the Web.

Acknowledgments: The authors would like to thank the two anonymous reviewers for their high quality and constructive input. The study conducted took place as part of the Advanced topics of Web Science seminar, organised by the Chair of Digital Libraries and Web Information Systems at the Faculty of Informatics and Mathematics, University of Passau.

Author Contributions: The authors contributed equally to this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

WWW	World Wide Web
W3C	World Wide Web Consortium
IETF	Internet Engineering Task Force
AFRINIC	African Network Information Centre
DMOZ	directory.mozilla.org

References

1. Rhodenizer, D.; Trudel, A. *How Big is the World Wide Web?* ICWI: Kingston, Jamaica, 2002; pp. 176–183.
2. Van den Bosch, A.; Bogers, T.; De Kunder, M. Estimating search engine index size variability: A 9-year longitudinal study. *Scientometrics* **2016**, *107*, 839–856.
3. Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; Wiener, J. Graph structure in the web. *Comput. Netw.* **2000**, *33*, 309–320.
4. Gulli, A.; Signorini, A. Building an open source meta-search engine. In Proceedings of the Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, Chiba, Japan, 10–14 May 2005; ACM: New York, NY, USA, 2005; pp. 1004–1005.
5. Brin, S.; Page, L. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw.* **2012**, *56*, 3825–3833.
6. Björneborn, L.; Ingwersen, P. Toward a basic framework for webometrics. *J. Assoc. Inf. Sci. Technol.* **2004**, *55*, 1216–1227.
7. Björneborn, L.; Ingwersen, P. Perspective of webometrics. *Scientometrics* **2001**, *50*, 65–82.
8. Spink, A.; Jansen, B.J.; Blakely, C.; Koshman, S. A study of results overlap and uniqueness among major web search engines. *Inf. Process. Manag.* **2006**, *42*, 1379–1391.
9. Taneja, H. Mapping an audience-centric World Wide Web: A departure from hyperlink analysis. *New Media Soc.* **2016**, *19*, 1331–1348.
10. Bharat, K.; Broder, A. A technique for measuring the relative size and overlap of public web search engines. *Comput. Netw. ISDN Syst.* **1998**, *30*, 379–388.
11. Kleinberg, J.; Kumar, R.; Raghavan, P.; Rajagopalan, S.; Tomkins, A. The web as a graph: Measurements, models, and methods. In Proceedings of the International Computing and Combinatorics Conference, Tokyo, Japan, 26–28 July 1999; pp. 1–17.
12. Albert, R.; Jeong, H.; Barabási, A.L. Internet: Diameter of the world-wide web. *Nature* **1999**, *401*, 130–131.
13. Orduña-Malea, E.; Ayllón, J.M.; Martín-Martín, A.; López-Cózar, E.D. Methods for estimating the size of Google Scholar. *Scientometrics* **2015**, *104*, 931–949.
14. Khabsa, M.; Giles, C.L. The number of scholarly documents on the public web. *PLoS ONE* **2014**, *9*, e93949.
15. Greenfield, D.N.; Davis, R.A. Lost in cyberspace: The web@work. *CyberPsychol. Behav.* **2002**, *5*, 347–353.
16. Bar-Ilan, J. Search engine results over time: A case study on search engine stability. *Cybermetrics* **1999**, *2*, 1–16.
17. Spink, A.; Jansen, B.J.; Kathuria, V.; Koshman, S. Overlap among major web search engines. *Internet Res.* **2006**, *16*, 419–426.
18. Gulli, A.; Signorini, A. The indexable web is more than 11.5 billion pages. In Proceedings of the Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, Chiba, Japan, 10–14 May 2005; ACM: New York, NY, USA, 2005; pp. 902–903.
19. Xing, S.; Paris, B.P. Measuring the size of the Internet via importance sampling. *IEEE J. Sel. Areas Commun.* **2003**, *21*, 922–933.
20. Lewandowski, D.; Wahlig, H.; Meyer-Bautor, G. The freshness of web search engine databases. *J. Inf. Sci.* **2006**, *32*, 131–148.
21. Lewandowski, D. A three-year study on the freshness of web search engine databases. *J. Inf. Sci.* **2008**, *34*, 817–831.
22. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55.