A Longitudinal Analysis of Search Engine Index Size

Antal van den Bosch[1], Toine Bogers[2], and Maurice de Kunder[3]

[1] *a.vandenbosch@let.ru.nl*
Centre for Language studies, Radboud University, PO Box 9103, 6500 HD Nijmegen (The Netherlands)

[2] *toine@hum.aau.dk*
Department of Communication, Aalborg University Copenhagen, A.C. Meyers Vænge 15, 2450 Copenhagen (Denmark)

[3] *maurice@dekunder.nl*
De Kunder Internet Media, Toernooiveld 100, 6525 EC, Nijmegen (The Netherlands)

Abstract
One of the determining factors of the quality of Web search engines is the size of their index. In addition to its influence on search result quality, the size of the indexed Web can also tell us something about which parts of the WWW are directly accessible to the everyday user. We propose a novel method of estimating the size of a Web search engine's index by extrapolating from document frequencies of words observed in a large static corpus of Web pages. In addition, we provide a unique longitudinal perspective on the size of Google and Bing's indexes over a nine-year period, from March 2006 until January 2015. We find that index size estimates of these two search engines tend to vary dramatically over time, with Google generally possessing a larger index than Bing. This result raises doubts about the reliability of previous one-off estimates of the size of the indexed Web. We find that much, if not all of this variability can be explained by changes in the indexing and ranking infrastructure of Google and Bing. This casts further doubt on whether Web search engines can be used reliably for cross-sectional webometric studies.

Conference Topic
Webometrics

Introduction
Webometrics (or cybermetrics) is commonly defined as the study of the content, structure, and technologies of the World Wide Web (WWW) using primarily quantitative methods. Since its original conception in 1997 by Almind & Ingwersen, researchers in the field have studied aspects such as the link structure of the WWW, credibility of Web pages, Web citation analysis, the demographics of its users, and search engines (Thelwall, 2009). The size of the WWW, another popular object of study, has typically been hard to estimate, because only a subset of all Web pages is accessible through search engines or by using Web crawling software. Studies that attempt to estimate the size of the WWW tend to focus on the surface Web—the part indexed by Web search engines—and often only at a specific point in time.

In the early days of search engines, having the biggest index size provided search engines with a competitive advantage, but a changing focus on other aspects of search result quality, such as recency and personalization, has diminished the importance of index size in recent years. Nevertheless, the size of a search engine's index is important for the quality of Web search engines, as argued by Lewandowski and Höchstötter (2008). In addition, knowledge of the size of the indexed Web is important for webometrics in general, as it gives us a ceiling estimate of the size of the WWW that is accessible by the average Internet user.

The importance of index sizes in the early days of Web search resulted in several estimation methods, most of which used the overlap between different Web search engines to estimate the size of the indexed Web as a whole. Bharat and Broder (1998) used an overlap-based method to estimate the size of the WWW at around 200 million pages. Lawrence & Giles

(1998, 1999) produced higher estimates of 320 and 800 million pages in 1998 and 1999 using a similar method, and Gulli and Signorini (2005) updated these estimates to 11.5 billion pages. The last decade has seen little work on index size estimation, but a general problem with all of the related work so far is that all the analyses have been cross-sectional. There has been no analysis of index size on a longer time scale that sheds light on the robustness of the different estimation methods. The handful of studies that have taken a longer-term perspective have typically focused on Web page persistence (Koehler, 2004) or academic link structure (Payne & Thelwall, 2008), but never search engine index size.

In this paper we present a novel method of estimating the size of a Web search engine's index by extrapolating from document frequencies of words observed in a large static corpus of Web pages. In addition, we provide a unique longitudinal perspective on our estimation method by applying it to estimate the size of Google and Bing's[1] indexes over a period of close to nine years, from March 2006 until January 2015.[2]

We find that index size estimates of these two search engines tend to vary wildly over time, with Google generally possessing a larger index than Bing. This considerable variability has been noted in earlier work (e.g., Rousseau, 1999; Payne & Thelwall, 2008), which raises doubts about the reliability of previous one-off estimates of the size of the indexed Web. In our analysis, we find that much of this variability can be explained by changes in the indexing and ranking infrastructure of Google and Bing. This casts further doubt on whether Web search engines can be used reliably for one-off Webometric studies, confirming similar sentiments expressed by, for instance, Payne and Thelwall (2008), and Thelwall (2012).

The remainder of this paper is organized as follows. The next section contains a review of related work in webometrics and on estimating the size of the indexed WWW. We then explain our estimation method in more detail, followed by the results of our estimation method and an analysis of the variability we uncover. We then discuss our findings and draw our conclusions.

Related work

Since its inception, researchers have studied many different aspects of the Web. This section provides a brief overview of some of the key studies on measuring different properties of Web search engines and the WWW, in particular work on estimating their size.

Measuring the Web

Over the past two decades many aspects of the WWW have been studied, such as the link structure of the Web that emerges from the hyperlinks connecting individual Web pages. Broder et al. (2000) were among the first to map the link structure of the WWW. They showed that the Web graph can be visualized as a bow-tie structure with 90% of all pages being a part of the largest strongly connected component, which was confirmed in 2005 by Hirate et al. (2008). Payne and Thelwall (2008) performed a longitudinal analysis of hyperlinks on academic Web sites in the UK, Australia and New Zealand over a six-year period. They found that the inlink and outlink counts were relatively stable over time, albeit with large fluctuations at the individual university level. As a result, they concluded that such variability could create problems for the replicability and comparability of webometrics research. Other related work on analyzing the link structure of the Web includes Kleinberg et al. (1999) and Björneborn (2004).

[1] Formerly known as Microsoft Live Search until May 28, 2009.

[2] Recent daily estimates produced by our method can be accessed through http://www.worldwidewebsize.com/. The time series data displayed in Figure 1 are available online at http://toinebogers.com/?page_id=757.

Web search engines are an essential part of navigating the WWW and as a result have received much attention. Many different aspects of Web search have been investigated, such as ranking algorithms, evaluation, user behavior, and ethical and cultural perspectives. Bar-Ilan (2004) and Zimmer (2010) provide clear, multi-disciplinary overviews of the most important work on these aspects.

From a webometric perspective the hit counts, search engine rankings, and the persistence of the indexed URLs are highly relevant for the validity and reliability of webometric research using Web search engines. Rousseau (1999) was among the first to investigate the stability of search engine results by tracking the *hit counts*—the number of results indicated for a query—for three single-word search terms in Altavista and NorthernLight over a 12-week period in 1998. Altavista exhibited great variability over a longer time period, even with only three anecdotal query words. Rousseau attributed this to changes in Altavista's infrastructure with the launch of a new version in 1998. Thelwall (2008) also performed a cross-sectional, quantitative comparison of the hit counts and search engine results of Google, Yahoo!, and Live Search. He extracted 1,587 single-word queries from English-language blogs "based purely on word frequency criteria" (Thelwall, 2008, p. 1704), found strong correlations between the hit count estimates of all three search engines, and recommended using Google for obtaining accurate hit count estimates. Uyar (2009) extended Thelwall's work by including multi-word queries. He found that the number of words in the query significantly affects the accuracy of hit counts, with single-word queries providing nearly double the hit count accuracy as compared to multi-word queries. Finally, Thelwall and Sud (2012) investigated the usefulness of the Bing Search API 2.0 for performing webometric research. They examined, among other things, the hit count estimates and found that these can vary by up to 50% and should therefore be used with caution in webometric research.

Bar-Ilan et al. (2006) compared the rankings of three different Web search engines over a three-week period. They observed that the overlap in result lists for textual queries was much higher than for image queries, where the result lists of the different search engines showed almost no overlap. Spink et al. (2006) investigated the overlap between three major Web search engines based on the first results pages and found that 85% of all returned top 10 results are unique to that search engine.

The issue of Web page persistence in search engine indexes—how long does a Web page remain indexed and available—was first examined by Bar-Ilan (1999) for a single case-study query during a five-month period in 1998. She found that for some search engines up to 60% of the results had disappeared from the index at the end of the period. She hypothesized that the distributed nature of search engines may cause different results to be served up from different index shards at different points in time. Koehler (2004) reported on the results of a six-year longitudinal study on Web page persistence. He also provided an overview of different longitudinal studies on the topic and concluded, based on the relatively small number of studies that exist, that Web pages are not a particularly persistent medium, although there are meaningful differences between navigation and content pages.

Index size estimation

In the last two decades, various attempts have been directed at estimating the size of the indexed Web. Some approaches focus on estimating the index size of a single search engine directly, while a majority focuses on estimating the overlap to indirectly estimate the size of the total indexed Web.

Highly influential work on estimating index size was done by Bharat & Broder (1998), who calculated the relative sizes of search engines by selecting a random set of pages from one engine, and checking whether each page was indexed by another engine. They used 35,000 randomly generated queries of 6 to 8 words selected at random from a Web-based lexicon and

sent these queries to four search engines. One of every top-100 results pages was randomly selected, after which they calculated the relative sizes and overlaps of search engines by selecting this random set of pages from one engine, and checking whether the page was indexed by another engine. By combining their method with self-reported index sizes from the commercial search engines, they estimated the size of the WWW to be around 200 million pages. Gulli et al. (2005) extended the work of Bharat and Broder by increasing the number of submitted queries by an order of magnitude, and using 75 different languages. They calculated the overlap between Google, Yahoo!, MSN Live, and Ask.com, and updated the previous estimates to 11.5 billion pages in January 2005. Most approaches that use the work of Bharat and Broder as a starting point focus on improving the sampling of random Web pages, which can be problematic because not every page has the same probability of being sampled using Bharat and Broder's approach. Several researchers have proposed methods of near-uniform sampling that attempt to compensate for this ranking bias, such as Henzinger et al. (2000), Anagnostopoulos et al. (2006), and Bar-Yossef and Gurevich (2006, 2011).

Lawrence and Giles (1998) estimated the indexed overlap of six different search engines. They captured the queries issued by the employees of their own research institute and issued them to all six engines. The overlap among search engines was calculated on the aggregated result sets, after which they used publicly available size figures from the search engines to estimate the size of the indexed Web to be 320 million pages. Lawrence and Giles updated their previous estimates to 800 million Web pages in July 1999. Dobra et al. (2004) used statistical population estimation methods to improve upon the original 1998 estimate of Lawrence and Giles. They estimated that Lawrence and Giles were off by a factor of two and that the Web contained around 788 million Web pages in 1998. Khelghati et al. (2012) compared several of the aforementioned estimation methods as well as some proposed modifications to these methods. They found that a modified version of the approach proposed by Bar-Yossef et al. (2011) provided the best performance.

Estimating the Size of a Search Engine Index through Extrapolation

On the basis of a textual corpus that is fully available, both the number of documents and the term and document frequencies of individual terms can be counted. In the context of Web search engines, however, we only have reported hit counts (or document counts), and we are usually not informed about the total number of indexed documents. Since it is the latter we are interested in, we want to estimate the number of documents indexed by a search engine indirectly from the reported document counts.

We can base such estimates on a training corpus for which we have full information on document frequencies of words and on the total number of documents. From the training corpus we can extrapolate a size estimation of any other corpus for which document counts are given. Suppose that, for example, we collect a training corpus T of 500,000 web pages, i.e. $|T| = 500{,}000$. For all words w occurring on these pages we can count the number of documents they occur in, or their document count, $d_T(w)$. A frequent word such as *are* may occur in 250,000 of the documents, i.e., it occurs in about one out of every two documents; $d_T(are) = 250{,}000$. Now if the same word *are* is reported to occur in 1 million documents in another corpus C, i.e., its document count $d_C(are) = 1{,}000{,}000$, we can estimate by extrapolation that this corpus will contain about $|C| = \frac{d_C(are) \times |T|}{d_T(are)}$, i.e., 2 million documents.

There are two crucial requirements that would make this extrapolation sound. First, the training corpus would need to be representative of the corpus we want to estimate the size of. Second, the selection of words[3] that we use as the basis for extrapolation will need to be such

[3] We base our estimates on words rather than on multi-word queries based on the findings of Uyar (2009).

that the extrapolations based on their frequencies are statistically sound. We should not base our estimates on a small selection of words, or even a single word, as frequencies of both high-frequency and low-frequency words may differ significantly among corpora. Following the most basic statistical guidelines, it would be better to repeat this estimation for several words, e.g., twenty times, and average over the extrapolations.

A random selection of word types is likely to produce a selection with relatively low frequencies, as Zipf's second law predicts (Zipf, 1995). A well-known issue in corpus linguistics is that when any two corpora are different in genre or domain, very large differences are likely to occur in the two corpora's word frequencies and document frequencies, especially in the lower frequency bands of the term distributions. It is not uncommon that half of the word types in a corpus occur only once; many of these terms will not occur in another disjoint corpus, even if it is of the same type. This implies that extrapolations should not be based on a random selection of terms, many of which will have a low frequency of occurrence.

The selection of words should sample several high-frequency words but preferably also several other words with frequencies spread across the document frequency bands.

It should be noted that Zipf's law concerns word frequencies, not document frequencies. Words with a higher frequency tend to recur more than once in single documents. The higher the frequency of a word, the more its document frequency will be lower than its word frequency. A ceiling effect thus occurs with the most frequent words if the corpus contains documents of sufficient size: they tend to occur in nearly all documents, making their document frequencies about the same and approaching the actual number of documents in the corpus, while at the same time their word token frequencies still differ to the degree predicted by Zipf's law (Zipf, 1995). This fact is not problematic for our estimation goal, but it should be noted that this hinges on the assumption that the training corpus and the new corpus of which the frequencies are unknown, contain documents of about the same average size.

As our purpose is to estimate the size of a Web search engine's index, we must make sure that our training corpus is representative of the web, containing documents with a representative average size. This is quite an ambitious goal. We chose to generate a randomly filtered selection of 531,624 web pages from the DMOZ[4] web directory. We made this selection in the spring of 2006. To arrive at this selection, first a random selection was made of 761,817 DMOZ URLs, which were crawled. Besides non-existing pages, we also filtered out pages with frames, server redirects beyond two levels, and client redirects. In total, the DMOZ selection of 531,624 documents contains 254,094,395 word tokens (4,395,017 unique word types); the average DMOZ document contains 478 words.

We then selected a sequence of DMOZ words by their frequency rank, starting with the most frequent word, and selecting an exponential series where we increase the selection rank number with a low exponent, viz. 1.6. We ended up with a selection of the following 28 words, the first nine being high-frequency function words and auxiliary verbs: *and*, *of*, *to*, *for*, *on*, *are*, *was*, *can*, *do*, *people*, *very*, *show*, *photo*, *headlines*, *william*, *basketball*, *spread*, *nfl*, *preliminary*, *definite*, *psychologists*, *vielfalt*, *illini*, *chèque*, *accordée*, *reticular*, *rectificació*. The DMOZ directory is multilingual, but English dominates. It is not surprising that the tail of this list contains words from different languages.

Our estimation method then consists of retrieving document counts for all 28 words from the search engine we wish to estimate the number of documents for, obtaining an extrapolated estimate for each word, and averaging (taking a mean) over the 28 estimations. If a word is not reported to occur in any document (which hardly happens), it is not included in the average.

[4] DMOZ is also called the Open Directory Project, http://www.dmoz.org/.

To stress-test the assumption that the DMOZ document frequencies of our 28 words yield sensible estimates of corpus size, we estimated the size of a range of corpora: the New York Times part of the English Gigaword corpus[5] (newspaper articles), the Reuters RCV1 corpus[6] (newswire articles), the English Wikipedia[7] (encyclopedic articles, excluding pages that redirect or disambiguate), and a held-out sample of random DMOZ pages. Table 1 provides an overview of the estimations on these widely different corpora. The size of the New York Times corpus is overestimated by a large margin of 126%, while the sizes of the other three corpora are underestimated. The size of the DMOZ sample—not overlapping with the training set, but drawn from the same source—is relatively accurately estimated with a small underestimation of 1.3%. Larger underestimations, for Reuters RCV1 and Wikipedia, may be explained by the fact that these corpora have shorter documents on average.

The standard deviations in Table 1, computed over the 28 words, indicate that the different estimates are dispersed over quite a large range. There seems to be no correlation with the size of the difference between the actual and the estimated number of documents. Yet, the best estimate, for the small DMOZ held-out sample (–1.3% error), coincides with the smallest standard deviation.

Table 1. Real versus estimated numbers (with standard deviations) of documents on four textual corpora, based on the DMOZ training corpus statistics: two news resources (top two) and two collections of web pages (bottom two). The second and third column provides the mean and median number of words per document.

	Words per document					
Corpus	*Mean*	*Median*	*# Documents*	*Estimate*	*St. dev.*	*Difference*
New York Times '94-'01	837	794	1,234,426	2,789,696	1,821,823	+126%
Reuters RCV1	295	229	453,844	422,271	409,648	– 7.0%
Wikipedia	447	210	2,112,923	2,024,792	1,385,105	– 4.2%
DMOZ test sample	477	309	19,966	19,699	5,839	– 1.3%

After having designed this experiment in March 2006, we started to run it on a daily basis on March 13, 2006, and have continued to do so. Each day we sent the 28 DMOZ words as queries to two search engines: Bing[i] and Google[8]. We retrieve the reported number of indexed pages on which each word occurs as it is returned by the web interface of both search engines, not their APIs. This number is typically rounded: it retains three or four significant numbers, the rest being padded by zeroes. For each word we use the reported document count to extrapolate an estimate of the search engine's size, and average over the extrapolations of all words. The web interfaces to the search engines have gone through some changes, and the time required to adapt to these changes sometimes caused lags of a number of days in our measurements. For Google 3,027 data points were logged, which is 93.6% of the 3,235 days between March 13, 2006 and January 20, 2015. For Bing, this percentage is 92.8% (3,002 data points).

Results

Figure 1 displays the estimated sizes of the Google and Bing indices between March 2006 and January 2015. For visualization purposes and to avoid clutter, the numbers are unweighted

[5] https://catalog.ldc.upenn.edu/LDC2003T05.
[6] http://trec.nist.gov/data/reuters/reuters.html.
[7] Downloaded on October 28, 2007.
[8] We also sent the same 28 words to two other search engines that were discontinued at some point after 2006.

running averages of 31 days, taking 15 days before and after each focus day as a window. The final point in our measurements is January 20, 2015; hence the last point in this graph is January 5, 2015. Rather than a linear, monotonic development we observe a rather varying landscape, with Google usually yielding the larger estimates. The largest peak in the Google index estimates is about 49.4 billion documents, measured in mid-December 2011. Occasionally, estimates are as low as under 2 billion pages (e.g. 1.96 billion pages in the Google index on November 24, 2014), but such troughs in the graph are usually short-lived, and followed by a return to high numbers (e.g., to 45.7 billion pages in the Google index on January 5, 2015).

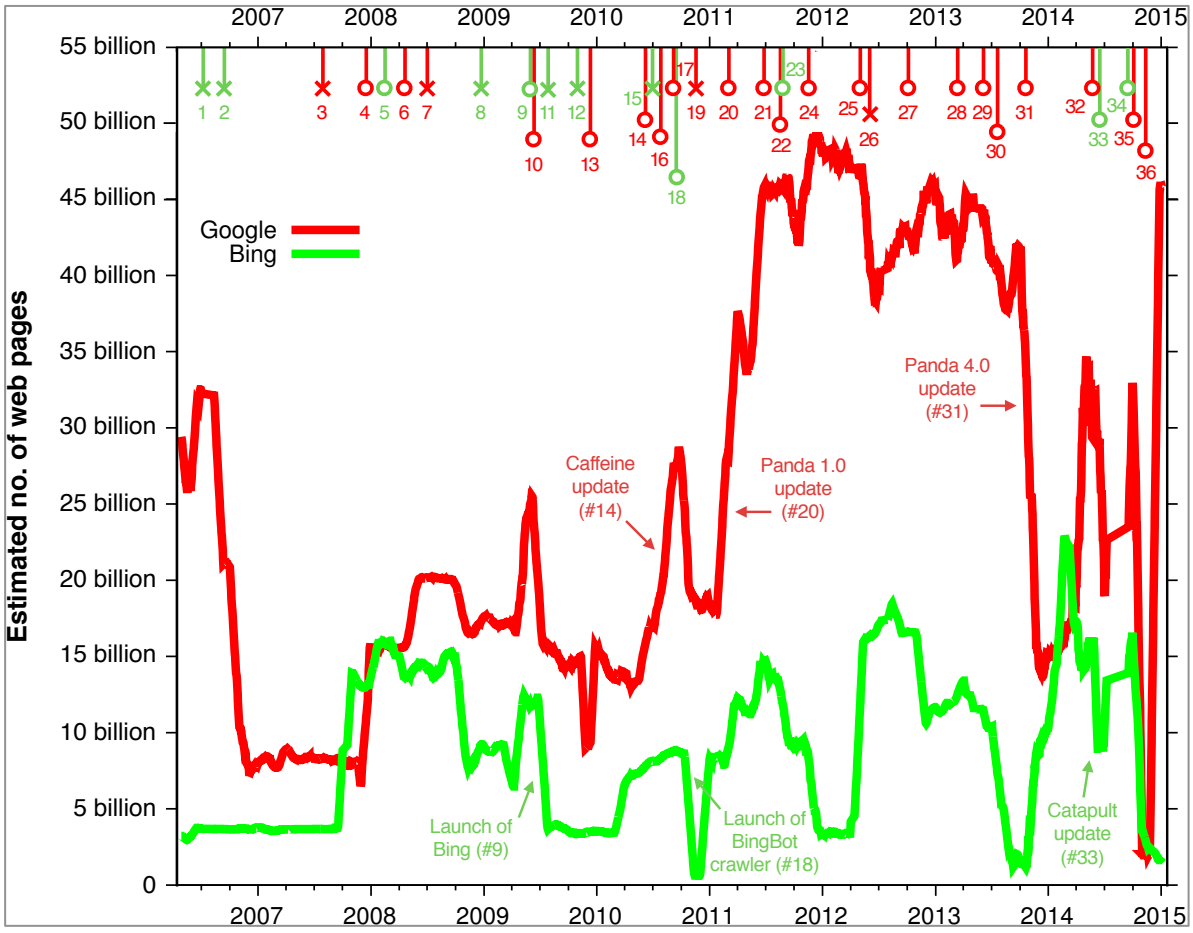

Figure 1. Estimated size of the Google and Bing indexes from March 2006 to January 2015. The lines connect the unweighted running daily averages of 31 days. The colored, numbered markers at the top represent reported changes in Google and Bing's infrastructure. The colors of the markers correspond to the color of the search engine curve they related to; for example, red markers signal changes in Google's infrastructure (the red curve). Events that line up with a spike are marked with an 'O', other events are marked with an '✕'.

Extrinsic variability

The variability observed in Figure 1 is not surprising given the fact that the indexing and ranking architectures of Web search engines are updated and upgraded frequently. According to Matt Cutts[9], Google makes "roughly 500 changes to our search algorithm in a typical year", and this is likely the same for Bing. While most of these updates are not publicized,

[9] http://googleblog.blogspot.com/2011/11/ten-algorithm-changes-on-inside-search.html.

some of the major changes that Google and Bing make to their architectures are announced on their official blogs. To examine which spikes in Figure 1 can be attributed to publicly announced architecture changes, we went through all blog posts on the Google Webmaster Central Blog[10], the Google Official Blog[11], the Bing Blog[12], and Search Engine Watch[13] for reported changes to their infrastructure. This resulted in a total of 36 announcements related to changes in the indexing or ranking architecture of Google and Bing[14]. The colored, numbered markers at the top of Figure 1 show how these reported changes are distributed over time.

For Google 20 out of the 24 reported changes appear to correspond to sudden spikes in the estimated index size, and for Bing 6 out of 12 reported changes match up with estimation spikes. This strongly supports the idea that much of the variability can be attributed to such changes. Examples include the launch of Bing on May 28, 2009 (event #9), the launch of Google's search index Caffeine on June 8, 2010 (event #14), the launch of the BingBot crawler (event #18), and the launches of Google Panda updates, and Bing's Catapult update (events #20, #31, and #33).

Of course not all spikes can be explained by reported events. For example, the spike in Bing's index size in October 2014 does not match up with any publicly announced changes in their architecture, although it is a likely explanation for such a significant change. In addition, some changes to search engine architectures are rolled out gradually and would therefore not translate to spikes in the estimated size. However, much of the variation in hit counts, and therefore estimated index size, appears to be caused by changes in the search engine architecture—something already suggested by Rousseau in his 1999 study.

Discussion and Conclusions

In this paper we presented a method for estimating the size of a Web search engine's index. Based on the hit counts reported by two search engines, Google and Bing, for a set of 28 words, the size of the index of each engine is extrapolated. We repeated this procedure and performed it once per day, starting in March 2006. The results do not show a steady, monotonic growth, but rather a highly variable estimated index size. The larger estimated index of the two, the one from Google, attains high peaks of close to 50 billion web pages, but occasionally drops to small indices of 2 billion pages as well. Are we measuring the extrinsic variability of the indices, or an intrinsic variability of our method? Our method is fixed: the same 28 words are sent to both search engines on every day. The frequencies of our test words are unlikely to change dramatically in a corpus as big as a crawl of the indexed Web; especially the document counts for our high-frequent words in our list should approximate (or at least be in the same order of magnitude as) the total number of documents in the index. We therefore believe that the variability we measure is largely, if not entirely attributable to the variability of the index of Google and Bing. In other words, what we are measuring is the genuine extrinsic variability of the indices, caused by changes (e.g., updates, upgrades, overhauls) of the indices. In Figure 1 we highlighted several publicly announced changes to both search engines' indices, many of which co-occur with drastic changes in index size as estimated by our method (20 out of the 24 reported changes in the Google index, and 6 out of 12 changes in Bing's index).

This variability, noted earlier also by Rousseau (1999), Bar-Ilan (1999), and Payne and Thelwall (2008), should be a cause for concern for any non-longitudinal study that adopts

[10] http://googlewebmastercentral.blogspot.com/.
[11] http://googleblog.blogspot.com/.
[12] http://blogs.bing.com/.
[13] http://searchenginewatch.com.
[14] A complete, numbered list of these events can be found at http://toinebogers.com/?page_id=757.

reported hit counts. It has been pointed out that "Googleology is bad science" (Kilgariff, 2007), meaning that commercial search engines exhibit variations in their functioning that do not naturally link to the corpus they claim to index. Indeed, it is highly unlikely that the real indexable Web suddenly increased from 20 to 30 billion pages in a matter of weeks in October 2014; yet, both the Bing and Google indices report a peak in that period. It is important to note, however, that the observed instability of hit counts does not automatically imply that measuring other properties of search engines for use in webometric research, such as result rankings or link structure, suffer from the same problem.

Our estimates do not show a monotonic growth of Web search engines' indices, which was one of the hypothesized outcomes at the onset of this study in 2006. The results could be taken to indicate that the indexed Web is not growing steadily the way it did in the late 1990s. They may even be taken to indicate the indexed Web is not growing at all. Part of this may relate to the growth of the unindexed Deep Web, and a move of certain content from the indexed to the Deep Web.

The unique perspective of our study is its longitude. Already in 1999, Rousseau remarked that collecting time series estimates should be an essential part of Internet research. The nine-year view visualized in Figure 1 shows that our estimation is highly variable. It is likely that other estimation approaches, e.g. using link structure or result rankings, would show similar variance if they were carried out longitudinally. Future work should include comparing the different estimation methods over time periods, at least of a few years. The sustainability of this experiment is non-trivial and should be planned carefully, including a continuous monitoring of the proper functioning. The scripts that ran our experiment for nearly nine years, and are still running, had to be adapted to changes in the web interfaces of Google and Bing repeatedly. The time required for adapting the scripts after the detection of a change caused the loss of 6-7% of all possible daily measurements.

Our approach, but also the different approaches discussed in the section on related research introduce different kinds of biases. We list here a number of possible biases and how they apply to our own approach:

Query bias. According to Bharat and Broder (1998), large, content-rich documents have a better chance of matching a query. Since our method of absolute size estimation relies on the hit counts returned by the search engines, it does not suffer from this bias, as the result pages themselves are not used.

Estimation bias. Our approach relies on search engines accurately reporting the genuine document frequencies of all query terms. However, modern search engines tend to not report the actual frequency, but instead estimate these counts, for several reasons. One such reason is their use of federated indices: a search engine's index is too large to be stored on one single server, so the index is typically divided over many different servers. Update lag or heavy load of some servers might prevent a search engine from being able to report accurate, up-to-date term counts. Another reason for inaccurate counts is that modern search engines tend to use document-at-a-time (DAAT) processing instead of term-at-a-time (TAAT) processing (Turtle & Flood, 1995). In TAAT processing the entire postings list is traversed for each query term in its entirety, disregarding relevant documents with each new trip down the postings list. In contrast, DAAT processing the postings list is traversed one document at a time for all query terms in parallel. As soon as a fixed number of relevant documents—say 1,000—are found, the traversal is stopped and the resulting relevant documents are returned to the user. The postings list is statically ranked before traversal (using measures such as PageRank) to ensure high quality relevant documents. Since DAAT ensures that, usually, the entire postings list does not have to be

traversed, the term frequency counts tend to be incomplete. Therefore, the term frequencies are typically estimated from the section of the postings list that was traversed.

Malicious bias. According to Bharat and Broder (1998, p. 384), "a search engine might rarely or never serve pages that other engines have, thus completely sabotaging our approach". This unlikely scenario is not likely to influence our approach negatively. However, if search engines were to maliciously inflate the query term counts, this would seriously influence our method of estimating the absolute index sizes.

Domain bias. By using text corpora from a different domain to estimate the absolute index sizes, a domain bias can be introduced. Because of different terminology, term statistics collected from a corpus of newswire, for instance, would not be applicable for estimating term statistics in a corpus of plays by William Shakespeare or corpus of Web pages. We used a corpus of Web pages based on DMOZ, which should reduce the domain bias considerably. However, in general the pages that are added to DMOZ are of high quality, and are likely to have a higher-than-average PageRank, which might introduce some differences between our statistics and the ideal statistics.

Cut-off bias. Some search engines typically do not index all of the content of all web pages they crawl. Since representative information is often at the top of a page, partial indexing does not have adverse effect on search engine performance. However, this cut-off bias could affect our term estimation approach, since our training corpus contains the full texts for each document. Estimating term statistics from, say, the top 5 KB of a document can have a different effect than estimating the statistics from the entire document. Unfortunately, it is impractical to figure out what cut-off point the investigated search engines use so as to replicate this effect on our training corpus.

Quality bias. DMOZ represents a selection of exemplary, manually selected web pages, while it is obvious that the web at large is not of the same average quality. Herein lies a bias of our approach. Some aspects of the less representative parts of the web have been identified in other work. According to Fetterly et al. (2005), around 33% of all Web pages are duplicates of one another. In addition, in the past about 8% of the WWW was made up of spam pages (Fetterly et al., 2004). If this is all still the case, this would imply that over 40% of the Web does not show the quality nor the variation present in the DMOZ training corpus.

Language bias. Our selection of words from DMOZ are evenly spread over the frequency continuum and show that DMOZ is biased towards the English language, perhaps more than the World Wide Web at large. A bias towards English may imply an underestimation of the number of pages in other languages, such as Mandarin or Spanish.

This exploratory study opens up at least the following avenue for future research that we intend to pursue. We have tacitly assumed that a random selection of DMOZ pages represents "all languages". With the proper language identification tools, by which we can identify a proper DMOZ subset of pages in a particular language, our method allows to focus on that language. This may well produce an estimate of the number of pages available on the Web in that language. Estimations for Dutch produce numbers close to two billion Web pages[ii]. Knowing how much data is available for a particular language, based on a seed corpus, is relevant background information for language engineering research and development that uses the web as a corpus (Kilgariff & Grefenstette, 2003).

References

Almind, T.C. & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'Webometrics'. *Journal of Documentation*, *53*, 404–426.

Anagnostopoulos, A., Broder, A. & Carmel, D. (2006). Sampling search-engine results. In *Proceedings of WWW '06*, (pp. 397–429). New York, NY, USA: ACM Press.

Bar-Ilan, J. (1999). Search engine results over time: a case study on search engine stability. *Cybermetrics*, *2*, 1.

Bar-Ilan, J. (2004). The use of Web search engines in information science research. *Annual Review of Information Science and Technology*, *38*, 231–288.

Bar-Ilan, J., Mat-Hassan, M. & Levene, M. (2006). Methods for comparing rankings of search engine results. *Computer Networks*, *50*, 1448–1463.

Bar-Yossef, Z. & Gurevich, M. (2006). Random sampling from a search engine's index. In *Proceedings of WWW '06* (pp. 367–376). New York, NY, USA: ACM Press.

Bar-Yossef, Z. & Gurevich, M. (2011). Efficient search engine measurements. *ACM Transactions on the Web*, *5*, 1–48.

Bharat, K. & Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of WWW '98* (pp. 379–388). New York, NY, USA: ACM Press.

Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. PhD thesis, Royal School of Library and Information Science.

Dobra, A. & Fienberg, S.E. (2004). How large is the World Wide Web? In *Web Dynamics* (pp. 23–43). Berlin: Springer.

Fetterly, D., Manasse, M. & Najork, M. (2005). Detecting phrase-level duplication on the World Wide Web. In *Proceedings of SIGIR '05* (pp. 170–177). New York, NY, USA: ACM Press.

Fetterly, D., Manasse, M. & Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam Web pages. In *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004* (pp. 1–6).

Gulli, A. & Signorini, A. (2005). The indexable Web is more than 11.5 billion pages. In *Proceedings of WWW '05* (pp. 902–903). New York, NY, USA: ACM Press.

Henzinger, M., Heydon, A., Mitzenmacher, M. & Najork, M. (2000). On near-uniform URL sampling. *Computer Networks*, *33*, 295–308.

Hirate, Y., Kato, S. & Yamana, H. (2008). Web structure in 2005. In Aiello, W., Broder, A., Janssen, J. & Milios, E. (Eds.) *Algorithms and Models for the Web-Graph*, vol. 4936, Lecture Notes in Computer Science, (pp. 36–46). Berlin: Springer.

Khelghati, M., Hiemstra, D. & Van Keulen, M. (2012). Size estimation of non-cooperative data collections. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services* (pp. 239–246). New York, NY, USA: ACM Press.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on Web as corpus. *Computational Linguistics*, *29*.

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, *33*, 147–151.

Kleinberg, J.M., Kumari, R., Raghavan, P., Rajagopalan, S. & Tomkins, A.S. (1999). The Web as a Graph: Measurements, Models, and Methods. In *COCOON '99: Proceedings of the 5th Annual International Conference on Computing and Combinatorics* (pp. 1–17). Berlin: Springer.

Koehler, W. (2004). A longitudinal study of Web pages continued: a report after six years. *Information Research*, 9. http://www.informationr.net/ir/9-2/paper174.html.

Lawrence, S. & Giles, C.L. (1998). Searching the World Wide Web. *Science*, *280*, 98–100.

Lawrence, S. & Giles, C.L. (1999). Accessibility of Information on the Web. *Nature*, *400*, 107–109.

Lewandowski, D. & Höchstötter, N. (2008). Web searching: a quality measurement perspective. In Spink, A. & Zimmer, M. (Eds.) *Web Search*, vol. 14, Information Science and Knowledge Management (pp. 309–340). Berlin: Springer.

Payne, N. & Thelwall, M. (2008). Longitudinal trends in academic Web links. *Journal of Information Science*, *34*, 3–14.

Rousseau, R. (1999). Daily time Series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, *2*, 1.

Spink, A., Jansen, B.J., Kathuria, V. & Koshman, S. (2006). Overlap among major Web search engines. *Internet Research*, *16*, 419–426.

Thelwall, M. (2008). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, *59*, 1702–1710.

Thelwall, M. (2009). Introduction to Webometrics: Quantitative Web Research for the Social Sciences. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1, pp. 1–116.

Thelwall, M. & Sud, P. (2012). Webometric research with the Bing search API 2.0. *Journal of Informetrics*, *6*, 44–52.

Turtle, H. & Flood, J. (1995). Query evaluation: Strategies and optimizations. *Information Processing & Management*, *31*, 831–850.

Uyar, A. (2009). Investigation of the accuracy of search engine hit counts. *Journal of Information Science*, *35*, 469–480.

Zimmer, M. (2010). Web Search Studies: Multidisciplinary Perspectives on Web Search Engines. In Hunsinger, J., Klastrup, L. & Allen, M. (Eds.) *International Handbook of Internet Research* (pp. 507–521). Berlin: Springer.

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort*; Reading, MA: Addison-Wesley.