WS 2015 Project 1:
Web Size

Anonymous

1. INTRODUCTION

The internet has grown exponentialy since Lawrence and Gile's paper on estimating the size of the Web. Their research has evolved into different types of methods that are currently used to measure the coverage of the major World Wide Web search engines. Estimating the size of the whole Web is quite difficult, due to its dynamic nature. In 1998, Lawrence and Giles [6, 3] gave a lower bound 800 million pages. These estimates are now obsolete.

The methodology that Gulli and Signorini used to conduct a new measurement of the size of the web is very similar to Lawrence and Giles approach. Their approach was based on analyzing the top 3 search engines from 2005: Google (which has the largest number of indexed pages of any search engine), Yahoo!, Ask/Teoma and MSM. From their findings, Google indexed around 68.2% of any other search engine, MSN indexed around 49.2%, Ask/Teoma index around 43.5% and Yahoo! index about 59.1%. Averaging their values, they estimated the Indexable Web to be approximately 11.5 billion pages. As reported in their paper, the estimated intersection of all four indexes is 28.85%, or about 2.7 billion pages, and their union is about 9.36 billion pages.

Another way of crawling the web in search was presented by Brian and Alvin Moore in the "Sizing the Internet" paper published in 2000. Their study was done by using the Cyveillence proprietary technology (NetSapien Technology). Their approach was by analysing 350 million links and crawling each page for other distinct URLs[]. The model analyzes spesific data associated with the links. The study ran over a four month period. By continuously referencing pages and exmining the links, the model is able to track the frequency with which unique URLs are encountered, both for the first time and each time thereafter. Their finding had surpassed their expectations with the number of unique pages on the Internet of 2.1 billion, and number of unique pages added per day of 7.3 million.

Web Science 2015 DIKU, Denmark

2. METHODOLOGY

The methodology I used for estimating the size of the web is very similar to Giles and Lawrence's approach from 1998. For crawling the web I used 500 queries and 3 search engines. The search engines that were considered for this experiment are: Bing, Baidu and Yandex. I have used these ones because Google, Yahoo! and other bigger search engines have protection mechanisms that prevent artificial polling of search results. If the crawler request to many searches in a short amount of time, Google enforces a cool down period for additional queries. In 2014, Bing reports that it has indexed more than 17 billion pages [https://www.ventureharbour.com/visualising-size-google-bing-yahoo/s]. Baidu has indexed over 740 million web pages, 80 million images, and 10 million multimedia files. ["MSN Money âĂŞ BIDU". MSN Money. Archived from the original on May 1, 2006. Retrieved 2006-05-11.]. Yandex has around 15 billion indexed pages[Project description].

The software that I used for crawling these search engines is GoogleScraper[https://github.com/NikolaiT/GoogleScraper](A Python module to scrape several search engines (like Google, Yandex, Bing, Duckduckgo, Baidu and others) by using proxies (socks4/5, http proxy) and with many different IP's, including asynchronous networking support). It worked decent enough for carrying out this assignment and it has some extra features that helped improve performance during crawling (for example multithreading support). The experiment was carried out in a time span of 2 days of non-stop crawling and data gathering.

2.1 Crawling Particularities

For each of the 500 queries I retrieve 20 pages of results. Twenty pages seemed sufficient for retrieving relative search results, because search engines often retrieve similar results for queries and not the actual query term. Each page contains between 10-15 results. GoogleScraper scrapes the three search engines in the same time with 10 worker threads. GoogleScrapes saves the URLs in the google_scraper.db SQL database. After finishing all 30000 keywords we begin analyzing the newly created database of links, references, search terms and search engines to make an estimation on how big is the Web.

To create a realistic estimate of the size of the Web, based on the current parameters, we would be checking the unique links retrieved by the crawler. Some links are redirects to different pages (for example www.baidu.com/safiahsiohasda4124124) we will leave them out of the data analysis because we don not know if the links would take us to the same pages that other search engines have retrieved for us. Also, because we are comparing links relative to actual documents we do not know if the pages contain the same information themselves. This will have an impact on the estimate of our web size. In 'Searching the world wide web' they state that they download each page and check that the query term occurs in the document. This is a very high computational task and we would require different hardware in order to obtain the same results.

Unfortunately after crawling for two days, I realized that bing and Baidu have autocorrection by default and they would retrieve the same results for two different queries. For example the queries "apple" and "aple" will return the same search results. This omission greatly impacts our search results.

3. FINDINGS

After crawling roughly 27 000 pages across 3 search engines the GoogleScraper program had managed to retrieve 179875 URLs. Out of these URLs only 110 478 are usable in our experiment because the script had some problems when generating request using the 127.0.0.1 IP address. The requests sent by that IP address had all entries for the search engines NULL. Thus we cannot conduct an analyzing only the URL because we don not know from what search engine it came from. Out of 110 478 links only 65 565 are unique.

URLs	Scraped	Usable
Total URLs	179 875	110 478
Unique URLs	69 385	65 565

Table 1: Total and Unique URLs scraped

As we can see from Table 2 there are great discrepancies between the 3 search engines. Yandex only retrieves 465 URLs compared to the other two which have dominant numbers (roughly between 51 000 and 60 000 links). From comparing the numbers it seems that GoogleScraper does not work well when crawling with the Yandex search engine. Another problem seems to be with the links retrieved from Baidu. These links are not direct links to pages but instead they are assigned a token that is relevant only to Baidu's routing infrastructure. Thus, we cannot conduct a proper comparison on what links are duplicated between the 3 search engines. When intersecting the Yandex results with Bing's we get only 11 common links. This value is simply too small to draw a conclusion on how relevant the results are. Baidu's links are unique and are not relevant in the analysis.

Search Engine	**Bing**	**Baidu**	**Yandex**
Total URLs	58 130	51 883	465
Unique URLs	13 241	51 870	465

Table 2: URLs categorized by search engine

4. CONCLUSIONS

The experiment that I did turns out to be inconclusive. The data is not consistent enough to make an accurate analysis. The problem resides in both the scraping tool utilized and the search engines. The tool proves to be unstable enough to skew the results and the search engines implement different methods of handling URLs and traffic balancing so that a crawler cannot add overhead to these services. However, these results show express the necessity of a different approach when it comes to crawling the web. In a future experiment I would be inclined to use another scraping tool and different search engines. The approach of Lawrence and Gile might not be relevant anymore. Modern search engines tackle the problem of searching the web in an entirely different manner compared to roughly 17 years ago. This experiment does provide some insight on how to conduct a new, more accurate way of measuring the size of the Internet.

5. REFERENCES

[1] A. S. A. Gulli. The indexable web is more than 11.5 billion pages. *Science*, pages 902–903, 2005.