



Differences in Time Delay between Search Engine Crawlers at Web Sites

Jeeva Jose¹ and P. Sojan Lal²

¹Department of Computer Applications, ²School of Computer Sciences

¹BPC College, Piravom, Kerala, India

²Mahatma Gandhi University, Kottayam, Kerala, India

Abstract: Web log mining provides tremendous information about user traffic and search engine behavior at web sites. The behavior of search engines could be used in analyzing server load, quality of search engines, dynamics of search engine crawlers, ethics of search engines etc. Search engine crawlers are highly automated programs which are seldom regulated manually. These crawlers periodically visit the web sites to collect the information. The dynamicity of search engine crawlers could be identified with the time delay between two consecutive visits. The more the visits of a crawler to a web site, the more it contributes to the server load. We intend to see whether there is a significant difference in the time delay between visits of a search engine crawler. Similarly the time delay between visits of various search engine crawlers is also analyzed to identify the differences in their behavior.

Keywords: WWW; Web Logs; Search Engine Crawlers; Time Delay; Robots.

I. Introduction

Search engine crawlers also known as bots, spiders or robots play a vital role in indexing and updating the changes coming up in the web sites. They are highly automated programs which are never regulated manually [1][2]. The crawler is one of the important components of a search engine. The bots or crawlers periodically visit World Wide Web and updates the contents on the web. The log files maintained by web site administrators provide immense information about the user behavior and search engine traffic. Most of the works in web log mining is related to user behavior as it has applications in online sales, targeted advertising, online marketing, market basket analysis etc. There are open source software available like Google Analytics which measures the number of visitors, duration of the visits, the demographic from which the visitor comes etc. But it cannot identify search engine visits because Google Analytics track users with the help of JavaScripts and search engine crawlers do not enable the JavaScripts embedded in web pages when the crawlers visit the web sites [3].

The search engine crawlers initially access the robots.txt file which specifies the Robot Exclusion Protocol. Robots.txt is a text file kept at the root of the web site directory. The crawlers are supposed to access this file first before it crawls the web pages. The crawlers which access this file first and proceeds to crawling are known as ethical crawlers and other crawlers who do not access this file are called unethical crawlers. The robots.txt file contains the information about which pages are allowed for crawling and which all folders and pages are denied access. Certain pages and folders are denied access because they contain sensitive information which is not intended to be publically available. There may be situations where two or more versions of a page will be available one as html and other one as pdf. The crawlers can be made to avoid crawling the pdf version to avoid redundant crawling. Also certain files like JavaScripts, images, stylesheets etc can be avoided for saving the time and bandwidth. There are two ways to do this. First one is with the help of robots meta tag and the other one is with the help of robots.txt file. The robots.txt file contains the list of all user agents and the folders or pages which are disallowed [4]. The structure of a robots.txt file is follows.

User-agent:

Disallow:

“User-agent:” is the search engine crawler and “Disallow:” lists the files and directories to be excluded from indexing. In addition to “User-agent:” and “Disallow:” entries, comment lines are included by putting the # sign at the beginning of the line. For example all user agents are disallowed from accessing the folder /a.# All user agents are disallowed to see the /a directory.

User-agent: *

Disallow: /a/

Even though search engine crawlers are supposed to access the robots.txt file first certain crawlers do not access this file. The crawlers which initially access the robots.txt and then the subsequent files or folders are known as ethical crawlers whereas others are known as unethical crawlers. Some crawlers like “Googlebot”, “Yahoo! Slurp” and “MSNbot” cache the robots.txt file for a web site and hence during the modification of robots.txt file, these robots may disobey the rules. Certain crawlers avoid too much load on a server by crawling the server at a low speed during peak hours of the day and at a high speed during late night and early morning [2]. Recently web crawlers are used for focused crawling, shopbot implementation and value added services on the web. As a result more active robots are crawling on the web and many more are expected to follow which will increase the

search engine traffic and web server activity [5]. A large number of crawlers are available in the web and we intend to see whether there is a significant difference in the time delay between visits of search engine crawlers at web sites. The time delay between various search engine crawlers were also analyzed to see whether there is a significant difference exists in their behavior.

II. Background Literature

There are several works that mentions about the search engine crawler behavior. A forecasting model is proposed for the number of pages crawled by search engine crawlers at a web site [3]. Sun et al has conducted a large scale study of robots.txt [2]. A characterization study and metrics of search engine crawlers is done to analyse the qualitative features, periodicity of visits and the pervasiveness of visits to a web site [5]. The working of a search engine crawler is explained in [6]. Nielsen NetRatings is one of the leading internet and digital media audience information and analysis services. NetRatings have provided a study on the usage statistics of search engines in United States [7]. Commercial search engines play a lead role in World Wide Web information dissemination and access. The evidence and possible causes of search engine bias is also studied [8]. An empirical pilot study is done to see the relationship between JavaScript usage and web site usage. The intention was to establish whether JavaScript based hyperlinks attract or repel crawlers resulting in an increase or decrease in web site visibility [9]. The ethics of search engine crawlers is identified using quantitative models [10]. Analysis of the temporal behavior of search engine crawlers at web sites is also done [11]. In this work the differences in time delay between visits of search engine crawlers and among the search engine crawlers is studied. The results revealed that there is a significant difference in their behavior.

III. Methodology

A. Pre Processing of Web Log Files

Web log files need considerable amount of pre processing. The user traffic needs to be removed from this file as this work focuses on the search engine behaviour. Improper pre processing may bias the data mining tasks and lead to incorrect results. About 90% of the traffic generated at web sites is contributed by search engine crawlers [12]. The advantages of pre processing are

- The storage space is reduced as only the data relevant to web mining is stored.
- The user visits and image files are removed so that the precision of web mining is improved.

The web logs are unstructured and unformatted raw source of data. Unsuccessful status codes and entries pertaining to irrelevant data like JavaScripts, images, stylesheets etc including user information are removed. The most widely used log file formats are Common Log File Format and Extended Log File Format. The Common Log File format contains the following information: a) IP address b) authentication name c) the date-time stamp of the access d) the HTTP request e) the URL requested f) the response status g) the size of the requested file. The Extended Log File format contains additional fields like a) the referrer URL b) the browser and its version and c) the operating system or the user agent[13][14]. Usually there are three ways of HTTP requests namely GET, POST and HEAD. Most HTML files are served via GET method while most CGI functionality is served via POST or HEAD. The status code 200 is the successful status code [13].

Search engines are identified from their IP addresses and user agents used for accessing the web. The log file of a business organization www.nestgroup.net of 30 days ranging from May 1, 2011 to May 31, 2011 comprising of 31 days. Table 1 shows the results of pre processing. Those search engines whose number of visits less than 5 in a month is eliminated before further analysis. There were 13 distinct search engine crawlers. Certain search engine crawlers made several visits on one day itself where as some others made one or two visits a day. The prominent crawlers were Baiduspider, Bingbot, Discobot, Ezooms, Feedfetcher-Google, Googlebot, Gosospider, Ichiro, MJ12bot, MSNbot, Slurp, Sogou, Sosospider and Yandex. Some crawlers were not significant because they made less than 5 visits a month. It includes Alexa, Exabot, Magpie and Yrspider.

Table 1. Results of pre processing

Total number of records	2,68,858
Number of successful search engine requests	21,230
Number of distinct search engine crawlers	17
Number of search engine crawlers after pre processing	13
Number of visits chosen	100

The crawler Alexa is an ethical robot which initially accesses the robots.txt file. The Alexa crawler identifies itself as `ia_archiver` in the HTTP "User-agent" header field. It uses a World Wide Web crawl strategy. Basically, it starts with a list of known URLs from across the entire internet, then it fetches local links found as it goes. There are several advantages to this approach, most importantly that it creates the least possible disruption to the sites being crawled [15]. Baiduspider is the user agent of the search engine Baidu. It is a chinese search engine crawler which crawls the server depending on the server load. Baidu has several user agents like Baiduspider for web search, Baiduspider-mobile for mobile search, Baiduspider-image for image search, Baiduspider-video for video search, Baiduspider-news for news search, Baiduspider-favo for bookmark search and Baiduspider-ads for business search [16]. Bingbot is the crawler for bing search engine. It was developed by Microsoft. Earlier it was msnbot which performed crawling activities for bing but was replaced by bingbot in 2010 [17]. Discobot is the experimental web crawler for discovery engine. They are still crawling, and their web-site is still just an empty shell providing no information. A private alpha version of Discovery Engine became available in 2010. A beta version was released in 2011 [18]. Ezooms bot is from Ezooms.com which obtains content for unknown purpose. Ezooms bot uses the following user agent string `Ezooms Mozilla/5.0 (compatible;Ezooms/1.0;ezooms.bot@gmail.com)` [19].

Feedfetcher Google is a crawler from Google to keep up with new contents on the web. Google collects atom feeds and RSS feeds when users choose to add them to their Google homepage or Google Reader. Feedfetcher collects and periodically refresh these user-initiated feeds, but does not index them in Blog Search or Google's other search services [20]. Googlebot is a web crawling spider from Google. Googlebot uses huge set of computers to crawl billions of pages on the web. It uses an algorithmic process which involves computer programs to determine which sites to crawl, how often, and how many pages to fetch from each site. Googlebot's crawl process begins with a list of webpage URLs, generated from previous crawl processes and augmented with sitemap data provided by webmasters. As Googlebot visits each of these websites it detects links SRC and HREF on each page and adds them to its list of pages to crawl. New sites, changes to existing sites, and dead links are noted and used to update the Google index. Usually on an average Googlebot access the site not more than once every few seconds. However, due to network delays, it is possible that the rate will appear to be slightly higher over short periods. In general, Googlebot download only one copy of each page at a time. If Googlebot is downloading a page multiple times, it is probably because the crawler was stopped and restarted. Googlebot was designed to be distributed on several machines to improve performance and scale as the web grows. To reduce the bandwidth usage, many crawlers on machines located near the sites are sent. Therefore, the logs may show visits from several machines at google.com, all with the user-agent Googlebot [21]. Ichiro is a Japanese web spider sent by the search engine goo. MJ12bot is the search engine crawler from the UK based search engine Majestic-12. Majestic-12 operates a greatly enhanced crawl, with updates on its web scale back links index on a daily basis. This back links index is open for queries using a dedicated, high performance search at MajesticSEO.com. Majestic-12 continues to offer webmasters the ability to download data for their own sites for free via MajesticSEO, and continues to invest in the improvement of its crawler and search infrastructure [22]. MSNbot is a crawler developed by Microsoft for MSN search engine. MSN search engine offers webmasters the ability to slow down the crawl rate to accommodate web server load issues. Websites that are small in terms of the number of pages and whose content is not regularly updated probably will never need to set crawl delay settings. This bot will automatically adjust its crawl rate to an appropriate level based on the content it finds with each pass. Larger sites that have a great many pages of content may need to be crawled more deeply and more often so that their latest content may be added into the index [23].

Slurp is the web crawler from Yahoo. The user agent for slurp is `Mozilla/5.0 (compatible; Yahoo! Slurp;)`The original developer of Slurp was Inktomi and later Yahoo acquired Inktomi [24]. Sogou is the crawler from the chinese search engine sogou. It can search text, images, music and maps. Sosospider is a chinese crawler from Soso.com. It is owned by Tencent Holdings Limited. Yandex is a Russian search engine [19].

B. Analysis of Variance and Duncan's Multiple Range Test

Analysis of Variance (ANOVA) is used to analyse the variations in the collection of data among groups and within groups. ANOVA conducted using a single factor is known as one way ANOVA [25]. In this data set we have chosen time delay in seconds for 100 visits ($N=100$) of 13 search engine crawlers for study. The single factor considered is the time elapsed between two consecutive visits of search engine crawlers.

Null hypothesis, H_0 : The means are equal.

Alternate hypothesis, H_1 : The means are not equal. Let **k**: the number of levels or groups in the experiment
N: total number of subjects in the experiment, **n**: number of subjects in each group, **T**: $\sum X$ for each group, **G**: $\sum X$ for the entire experiment ($\sum T = G$), the following Table 2 gives the formula summary for ANOVA.

B. Table 2. Formula Summary Table for ANOVA

Source	df	SS	MS	F	p
Between Groups	k-1	$\sum \frac{T^2}{n} - \frac{G^2}{N}$	$\frac{SS_{BG}}{df_{BG}}$	$\frac{MS_{BG}}{MS_{WG}}$	If $P > 0.10$ No evidence against the null hypothesis. If $0.05 < P < 0.10$, Weak evidence against the null hypothesis. If $0.01 < P < 0.05$ Moderate evidence against the null hypothesis. If $0.001 < P < 0.01$ Strong evidence against the null hypothesis. If $P < 0.001$ Very strong evidence against the null hypothesis.
Within Groups	N-k	$\sum SS$ inside each group	$\frac{SS_{WG}}{df_{WG}}$		
Total	N-1	$\sum X^2 - \frac{G^2}{N}$			

The detailed statistic descriptive of the time delay in seconds between visits of search engine crawlers is given in Table 3 and results of ANOVA is given in Table 4.

Table 3. Statistic Descriptives

Search Engine Crawlers	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Baiduspider	100	4440.00	5620.870	562.087	3324.70	5555.30	900	29100
Bingbot	100	3886.80	2811.607	281.161	3328.92	4444.68	60	18540
Discobot	100	8052.60	24497.378	2449.738	3191.79	12913.41	120	146220
Ezooms	100	15196.80	11557.235	1155.723	12903.59	17490.01	1260	57600
Feedfetcher-Google	100	1957.20	3282.790	328.279	1305.82	2608.58	480	24060
Googlebot	100	1488.00	1112.573	111.257	1267.24	1708.76	180	6000
Ichiro	100	2387.40	9988.802	998.880	405.40	4369.40	180	94620
MJ12bot	100	15843.00	9565.892	956.589	13944.92	17741.08	2100	43080
Slurp	100	3662.40	2097.304	209.730	3246.25	4078.55	120	9900
Sogou	100	13288.80	10824.459	1082.446	11140.99	15436.61	540	45660
Sosospider	100	16754.40	9455.527	945.553	14878.22	18630.58	600	41040
Yandex	100	5429.40	4455.277	445.528	4545.38	6313.42	180	33600
Yodao	100	8065.80	2669.603	266.960	7536.09	8595.51	4500	21840
Total	1300	7727.12	11045.041	306.334	7126.16	8328.09	60	146220

Table 4. One Way ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	38239422380.308	12	3186618531.692	34.111	.000
Within Groups	120229399260.000	1287	93418336.643		
Total	158468821640.308	1299			

The significance is .000 and we reject the null hypothesis and conclude that there is a significant difference in the mean time delay. Figure 1 shows the mean plot of the time delay of various search engine crawlers. Post hoc tests are used for situations in which the results have already obtained a significant omnibus F-test with a factor that consists of three or more means and additional exploration of the differences among means is needed to provide specific information on which means are significantly different from each other [26]. Table 5 shows the results of Duncan's multiple range test. The significant difference or the range value is given by

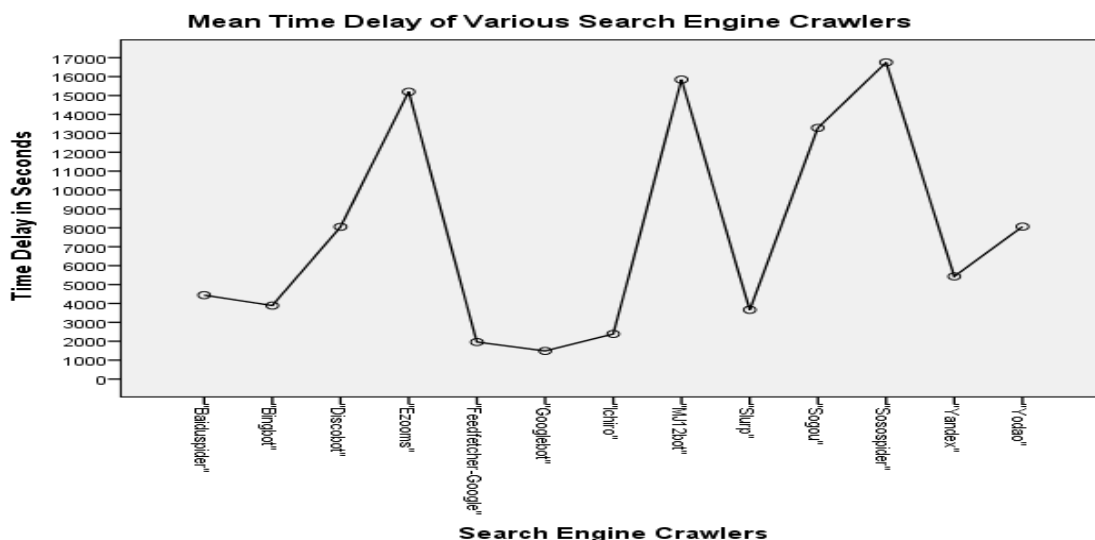
$$R_p = r_{\alpha,p,v} \sqrt{MSE/n} \quad (1)$$

where $r_{\alpha,p,v}$ is the Duncan's Significant Range Value with parameters p (= range-value) and v (= MSE degree-of-freedom), and experiment-wise alpha level α (= α_{joint}). MSE is the mean square error from the ANOVA table and n is the number of observations used to calculate the means being compared.

Table 5. Duncan's Multiple Range Test

Search Engine Crawlers	N	Subset for alpha = 0.05				
		1	2	3	4	5
Googlebot	100	1488.00				
Feedfetcher-Google	100	1957.20				
Ichiro	100	2387.40				
Slurp	100	3662.40	3662.40			
Bingbot	100	3886.80	3886.80			
Baiduspider	100	4440.00	4440.00			
Yandex	100		5429.40	5429.40		
Discobot	100			8052.60		
Yodao	100			8065.80		
Sogou	100				13288.80	
Ezooks	100				15196.80	15196.80
MJ12bot	100				15843.00	15843.00
Sosospider	100					16754.40
Sig.		.058	.244	.068	.077	.286

Figure 1. Mean plot of time delay of search engine crawlers



IV. Conclusion

The results of one way ANOVA showed that there is a significant difference in the time delay between search engine crawlers and among repeated visits of the same crawler. The mean plot and post hoc test revealed that the time delay of Googlebot, Feedfetcher-google and Ichiro was almost similar. Similarly Slurp, Bingbot and Baiduspider showed a similar time delay which was higher than that of the first set of crawlers including Googlebot. All other crawlers showed significant time delay in visiting the web site. This shows that certain crawlers like Googlebot, Feedfetcher-google, Ichiro, Slurp, Bingbot and Baiduspider are very dynamic compared to other crawlers like Ezzooms, MJ12bot, Sosospidar etc. Among the crawlers, Googlebot showed least time delay for repeated visits and exhibited consistency in its behavior. The time delay can be further studied to develop a forecasting model for the time delay of search engine crawlers.

References

- [1] C. Lee Giles, Yang Sun and Issac G. Council, "Measuring the Web Crawler Ethics," WWW2010, ACM, 2010, pp. 1101-1102.
- [2] Yang Sun, Ziming Zhuang and C. Lee Giles, "A Large- Scale Study of Robots.txt", WWW2007, ACM, 2007, pp.1123-1124.
- [3] Jeeva Jose, P. Sojan Lal, "A Forecasting Model for the Pages Crawled by Search Engine Crawlers at a Web Site", International Journal of Computer Applications(IJCA), Vol 68, Issue 13, 2013, pp.19-24.
- [4] <http://www.webconfs.com/what-is-robots-txt-article-12.php>, Accessed June 4, 2013.
- [5] Dikaikos M.P, Athena S. and Loizos P., "An Investigation of Web Crawler Behavior: Characterization and Metrics", Computer Communications, Vol 28, 2005, pp.880-897.
- [6] Brin .S and Page.L, The Anatomy of a Large Scale Hypertextual Web Search Engine, In *Proceedings of the 7th International WWW Conference*, Elsevier Science, New York, 1998.
- [7] Sullivan D., "Webspin: Newsletter " <http://contentmarketingpedia.com/Marketing-Library/Search/industryNewsSeptA1.pdf>, Accessed June 4, 2013.
- [8] Vaughan L. and Thelwal M., "Search Engine Coverage Bias: Evidence and Possible causes", Information Processing and Management, Vol 40, pp. 693-707.
- [9] Schwenke F. and Weideman M., "The Influence that JavaScript has on the visibility of a web site to search engines – a pilot study", Informatics & Design Papers and Reports, Vol 11, pp. 1-10.
- [10] C. Lee Giles, Yang Sun and Issac G. Council, "Measuring the Web Crawler Ethics," WWW2010, ACM, 2010, pp. 1101-1102.
- [11] Jeeva Jose, P. Sojan Lal, Analysis of the Temporal Behavior of Search Engine Crawlers at Web sites, COMPUSOFT, Vol. 2, Issue 6, 2013, pp.136-142.
- [12] D. Mican & D. Sitar-Taut, "Preprocessing and Content/ Navigational Pages Identification as Premises for an Extended Web Usage Mining Model Development", Informatica Economica, 2009, vol. 13(4), pp.168-179.
- [13] A. H. M. Wahab, H.N.M. Mohd, F.H. Hanaf & M.F.M. Mohsin, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science, Engineering and Technology, 2008, pp.190-197.
- [14] M. Spiliopoulou, "Web Usage Mining for Web Site Evaluation", Communications of the ACM, 2000. Vol. 43(8), pp.127-134.
- [15] <http://www.alexa.com/help/webmasters>, Accessed May10, 2013.
- [16] http://www.webmasterworld.com/search_engine_spiders/4348357.htm, Accessed May10, 2013.
- [17] <http://user-agent-string.info/list-of-ua/bot-detail?bot=bingbot>, Accessed May15, 2013.
- [18] <http://whatis.riskyinternet.com/what-is/web-robot/discoveryengine-robot-6142/> Accessed May10, 2013.
- [19] <http://www.rhyolite.com/anti-spam/badbots.html>, Accessed May10, 2013.
- [20] <http://support.google.com/webmasters/bin/answer.py?hl=en&answer=178852>, Accessed May10, 2013.
- [21] <http://support.google.com/webmasters/bin/answer.py?hl=en&answer=182072>, Accessed May15, 2013
- [22] <http://www.majestic12.co.uk/projects/dsearch/> Accessed May10, 2013
- [23] http://www.bing.com/blogs/site_blogs/b/webmaster/archive/2009/08/10/crawl-delay-and-the-bing-crawler-msnbot.aspx, Accessed May10, 2013
- [24] <http://help.yahoo.com/help/us/ysearch/slurp>, Accessed May 10, 2013
- [25] <http://www.stat.cmu.edu/~hseltman/309/Book/chapter7.pdf>, Accessed May 10, 2013
- [26] Duncan D. B., Multiple range and multiple F tests, Biometrics II, 1955, pp.1-42.

Acknowledgments

This research work is funded by Kerala State Council for Science Technology and Environment, Thiruvananthapuram, Kerala State, India as per order No.009/SRSPS/2011/CSTE.