ISSN:2320-0790

An International Journal of Advanced Computer Technology

Analysis of the Temporal Behaviour of Search Engine Crawlers at Web Sites

Jeeva Jose[1], P. Sojan Lal[2],

[1]Department of Computer Applications, BPC College, Piravom, India.
[2] School of Computer Science, Mahatma Gandhi University, Kottayam, India.

Abstract: Web log mining is the extraction of web logs to analyze user behaviour at web sites. In addition to user information, web logs provide immense information about search engine traffic and behaviour. Search engine crawlers are highly automated programs that periodically visit the web site to collect information. The behaviour of search engines could be used in analyzing server load, quality of search engines, dynamics of search engine crawlers, ethics of search engines etc. The time spent by various crawlers is significant in identifying the server load as major proportion of the server load is constituted by search engine crawlers. A temporal analysis of the search engine crawlers were done to identify their behaviour. It was found that there is a significant difference in the total time spent by various crawlers. The presence of search engine crawlers at web sites on hourly basis was also done to identify the dynamics of search engine crawlers at web sites.

Keywords: Web sites; Search engine; Crawlers; Web logs; Server Load;

I. INTRODUCTION

World Wide Web has seen a spectacular growth in terms of the number of websites and visitors since last decade. Without search engines crawlers, the web sites will not be visible to the users. 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][5]. 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 is 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 Java Scripts and search engine crawlers do not enable the Java Scripts 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 do avoid crawling the pdf version to avoid redundant crawling. Also certain files like Java Scripts, images, style sheets 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 [5]. 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 [6]. A large number of crawlers are available in the web and we intend to see whether there is a significant difference in the time spent by various crawlers at web sites. The time distribution of various crawlers over a period of 30 days on an hourly basis was analyzed to see the presence of search engine crawlers at web sites.

II. BACKGROUND LITERATURE

There are several works that mentions about the search engine crawler behaviour. 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 [5]. 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 [6]. The working of a search engine crawler is explained in [7]. Neilsen 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 [8]. 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 [9]. 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 [10]. The ethics of search engine crawlers is identified using quantitative models [11]. In this work search engine crawlers from two web sites are chosen for study to see the differences in their behaviour based on the total time spent and the distribution of time for each day.

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 files of 2 different organizations were selected for study. The first dataset is the log file of a business organization www.nestgroup.net of 30 days ranging from April 1, 2011 to April 30, 2011 and second dataset belongs to an academic website www.bpccollege.ac.in ranging from November 1, 2012 to November, 2012 comprising of 30 days. Table I shows the results of pre processing.

Table I. Results of pre processing

	Data Set 1	Data Set 2
Total number of records	2,65,476	1,45,680
Number of successful search engine requests	18,330	3,052
Number of distinct search engine crawlers	17	5

Those search engines whose number of visits less than 5 in a month is eliminated before further analysis. From data set 1 there were 14 distinct search engine crawlers and from data set 2 there were 2 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 time spent on a page is calculated by finding the difference between two consecutive requests. But it is difficult to predict when the crawler has left the last page. Table II and Table III shows various crawlers in data set 1 and data set 2 respectively with the total time spent in seconds for each day. The prominent crawlers in data set 1 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.

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. The 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].

There were five search engine crawlers in data set 2. But the prominent crawlers were Bingbot and Googlebot. The other crawlers were Ezooms, Yandex and Ahrefsbot. The Ahrefsbot is a fast SEO spy crawler originating from Ukraine. These bots wastes the bandwidth and are banned by many websites. The information collected by this bot is available for sale, provide opportunity for others to analyse one's content with the help of tools and also used for their own purpose also [25]. We intend to see whether there is a significant difference in the total time spent.

Table II. Total Time spent in seconds by various search engine crawlers in data set 1

	Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**Baiduspider**	0	4380	5880	5880	8580	10620	10620	55680	53880	44880	54180	37200	29460	12660	2400
2	**Bingbot**	11760	12420	13620	6660	11460	24360	24360	24480	28500	19200	16800	13320	8760	10980	17160
3	**Discobot**	0	36480	9600	12120	13020	17100	17100	6900	11040	6480	8700	0	2280	0	0
4	**Ezooms**	0	3540	3060	3960	3960	4140	4140	4200	6000	2640	2580	3900	3180	2520	1800
5	**Feedfetcher-Google**	17280	7680	8460	31140	32160	26760	26760	36000	5760	2760	24900	30840	26760	20400	20280
6	**Googlebot**	12900	16860	35100	20040	46800	37740	37740	23760	46560	10440	47340	59340	54780	49980	30300
7	**Gosospider**	960	780	3360	0	840	0	660	420	0	0	0	0	1740	1500	0
8	**Ichiro**	8460	28080	12480	10200	5700	8520	8520	20760	13560	13320	0	17040	24480	16140	16800
9	**MJ12bot**	480	60	180	120	180	60	1500	60	300	180	180	60	1440	180	240
#	**MSNbot**	0	1680	840	120	0	60	3780	0	60	0	0	0	0	3240	180
#	**Slurp**	13920	16440	10140	10320	19140	13740	13740	9180	10020	0	9480	23700	16320	9360	14520
#	**Sogou**	5580	1140	1680	0	4260	0	4620	2280	3000	0	1800	1620	0	360	300
#	**Sosospider**	0	0	2760	0	0	0	1320	0	0	1260	0	0	0	5160	0
#	**Yandex**	3420	3120	1200	900	1020	60	60	2520	6240	15000	12300	4500	3360	1080	5760

Table II Cont...

	Day	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	**Baiduspider**	1500	2220	2940	2760	3120	120	0	1020	2340	2040	2220	3420	420	2400	0
2	**Bingbot**	21000	11820	22440	18360	13440	8640	12000	6480	10320	12540	18600	8340	4680	5640	11760
3	**Discobot**	0	0	0	0	0	0	0	0	0	1860	1920	1500	0	0	0
4	**Ezooms**	1740	1800	2460	5700	2280	1140	1140	1500	1080	1080	2520	2340	1380	540	0
5	**Feedfetcher-Google**	14160	3780	26700	26220	27360	32160	16320	7860	6300	28680	36960	36840	35820	42240	17280
6	**Googlebot**	27600	34020	28740	48240	504600	40980	13140	46200	43800	28800	14160	35460	56400	21360	12900
7	**Gosospider**	1680	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	**Ichiro**	46860	23520	5640	4740	27060	20760	5040	26040	18060	3360	34440	7140	17100	41640	8460
9	**MJ12bot**	180	120	240	420	60	180	1380	1260	300	300	240	840	1800	1620	1020
10	**MSNbot**	1440	0	0	0	0	0	0	120	60	60	0	0	0	0	0
11	**Slurp**	17820	12120	10140	13200	11460	19140	12660	5880	7740	11520	14220	11580	13740	13620	13920
12	**Sogou**	1800	780	0	600	120	2760	1080	0	0	2220	360	0	2460	0	660
13	**Sosospider**	660	3600	0	2940	900	5700	0	0	1320	0	0	1500	0	0	0
14	**Yandex**	60	1500	6360	0	0	2760	3600	0	1680	3600	1800	5100	300	4380	1200

Table III. Total Time spent in seconds by various search engine crawlers in data set 2

Day	Bingbot	Googlebot	Day	Bingbot	Googlebot	Day	Bingbot	Googlebot
1	2040	3420	**11**	0	5520	**21**	0	3300
2	0	1740	**12**	0	2640	**22**	960	0
3	0	1620	**13**	0	2100	**23**	3240	11760
4	1620	4980	**14**	5760	4980	**24**	2220	2940
5	2340	4740	**15**	0	1740	**25**	0	5400
6	540	2640	**16**	1860	1020	**26**	11160	1140
7	2460	3000	**17**	10140	0	**27**	4380	900
8	3960	780	**18**	2940	6240	**28**	4560	1560
9	0	1500	**19**	0	3720	**29**	0	960
10	4020	2520	**20**	0	5460	**30**	0	3660

B. *Kruskal Wallis H Test*

Kruskal Wallis H Test detects if n data groups belong or not to the same population [26][27]. This statistic is a non parametric test suitable to distributions that are not normal such as the exponential distributions observed in web usage mining or web log analysis [28]. The formula for H static of Kruskal- Wallis test is given below where K is the number of samples.

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{\text{Rj2}}{\text{nj}} - 3(N+1) \quad (1)$$

Where Rj is the sum of the ranks of the sample j, nj is the size of the sample j, j=1, 2, 3, ...K and N is the size of the pooled sample (n_1+n_2+........n_K). The calculated H value is to be compared against the chi-square value with (K-1) degrees of freedom at the given significance level α.

H_0: There is no significant difference between the total time spent by various search engine crawlers.
H_1: There is significant difference between the total time spent by various search engine crawlers.
The test statistic for Kruskal Wallis H Test is shown in Table IV. For data set 1, the p-value shows a strong evidence of rejecting the null hypothesis and for data set 2 shows a moderate evidence of rejecting the null hypothesis. The result of H test shows that there is a significant difference in the total time spent by various search engines.

Table IV. Test Statistic

Kruskall Wallis Test		
	Data Set 1	Data Set 2
α	0.01	0.01
p-value	0.000	0.026
Chi-square	285.655	4.963
df	13	1

The time distribution of various crawlers in data set 1 and data set 2 were analysed for monitoring the presence of crawlers on an hourly basis in web sites. Figure 1 to Figure 14 shows the presence of prominent crawlers and their time distribution in data set 1. Figure 15 and Figure 16 shows the time distribution of crawlers in data set 2.

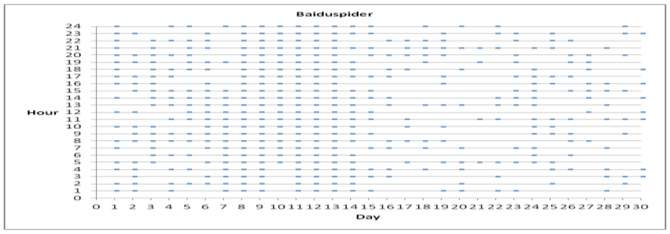

Figure 1. Time distribution for Baiduspider in Data Set 1

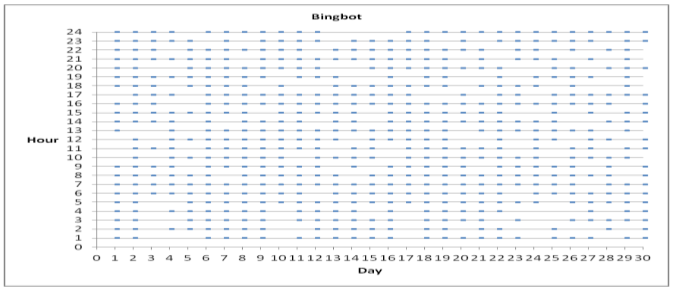

Figure 2. Time distribution for Bingbot in Data Set 1

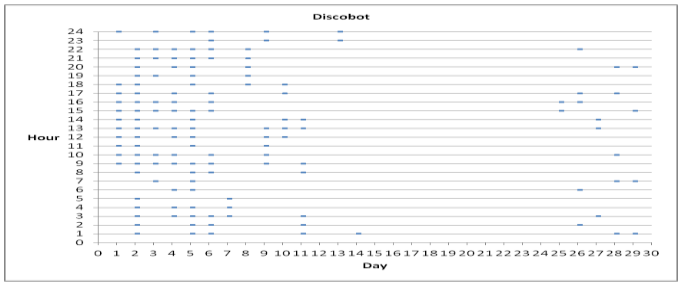

Figure 3. Time distribution for Discobot in Data Set 1

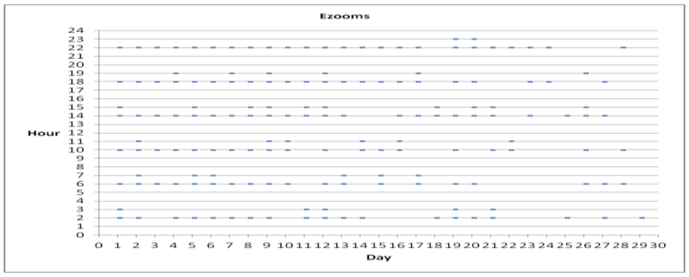

Figure 4. Time distribution for Ezooms in Data Set 1

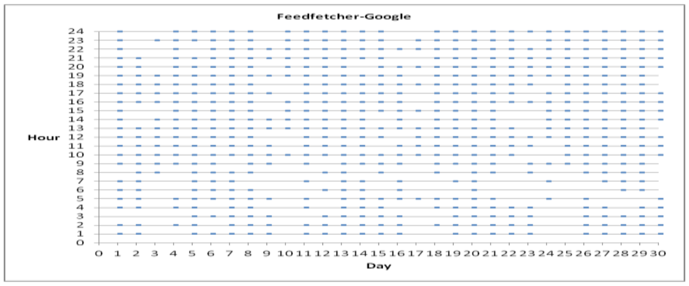

Figure 5. Time distribution for Feedfetcher-Google in Data Set 1

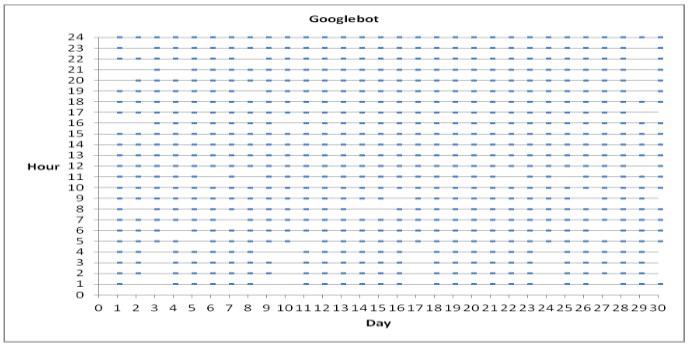

Figure 6. Time distribution for Googlebot in Data Set 1

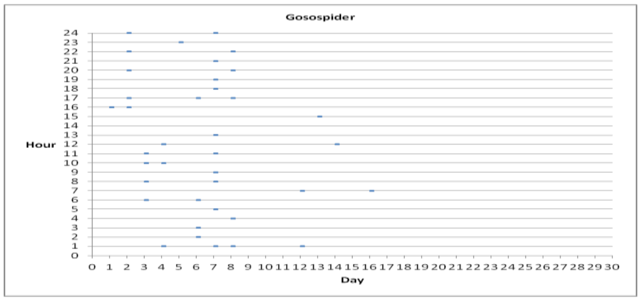

Figure 7. Time distribution for Gosospider in Data Set 1

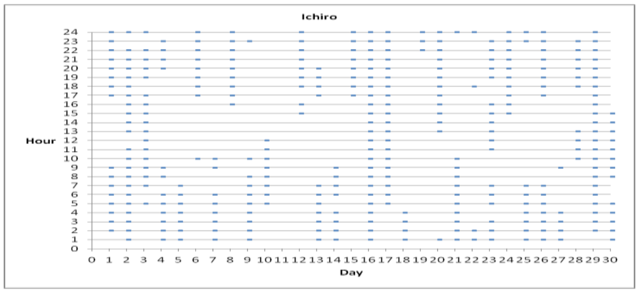

Figure 8. Time distribution for Ichiro in Data Set 1

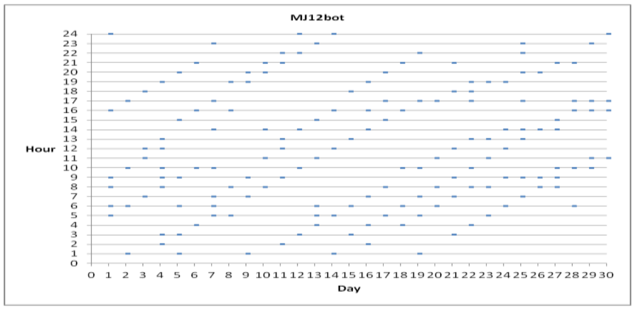

Figure 9. Time distribution for MJ12bot in Data Set 1

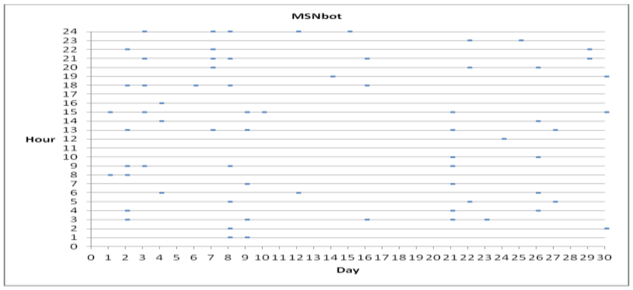

Figure 10. Time distribution for MSNbot in Data Set 1

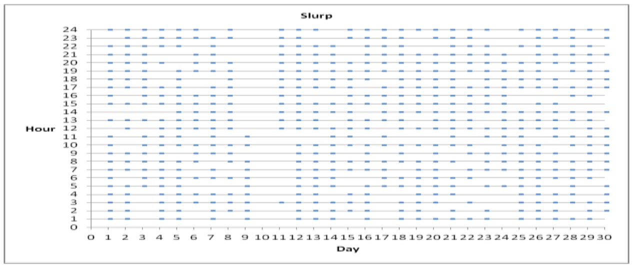

Figure 11. Time distribution for Slurp in Data Set 1

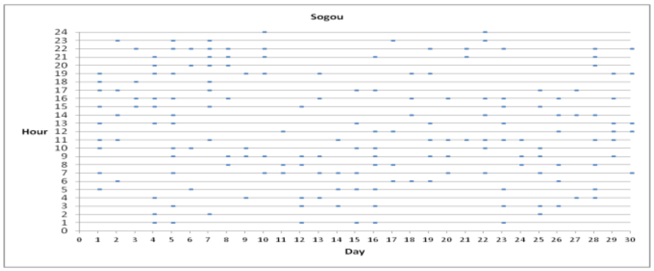

Figure 12. Time distribution for Sogou in Data Set 1

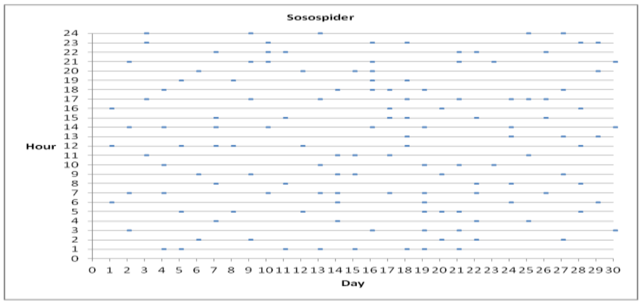

Figure 12. Time distribution for Sosospider in Data Set 1

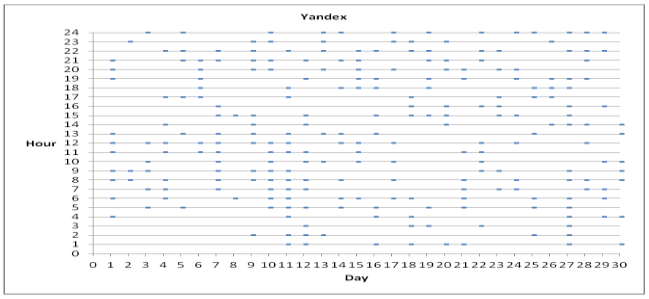

Figure 13. Time distribution for Yandex in Data Set 1

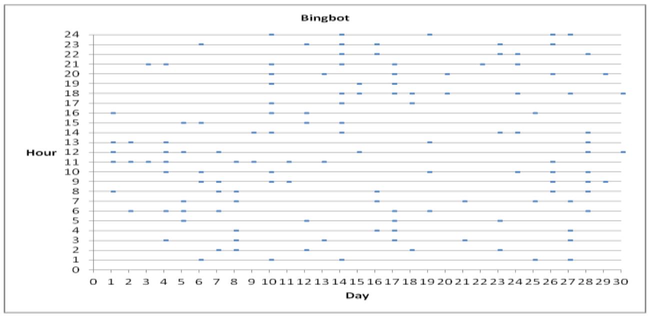

Figure 14. Time distribution for Bingbot in Data set 2

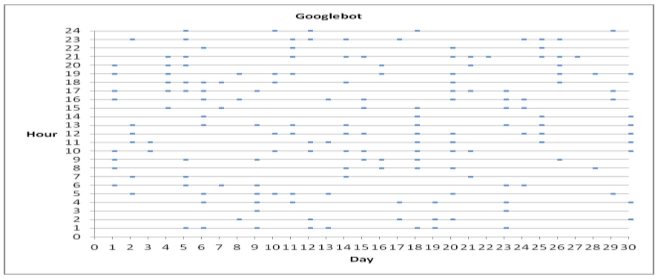

Figure 15. Time distribution for Googlebot in Data set 2

IV. Conclusion

The results of Kruskal Wallis H test showed that there is a significant difference in the behavior of search engine crawlers in terms of the total time spent for both data sets. Certain crawlers like Googlebot, Feedfetcher-Google, Bingbot, Baiduspider etc. showed consistency in the time spent whereas certain other bots like Gosospider, MSNbot, Discobot etc. were not consistent in their behavior. The crawlers like Googlebot, Feedfetcher-Google, Bingbot and Baiduspider were dynamic and present in almost every hour which contributes a major portion of the server load.

V. Acknowledgement

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.

VI. References

[1] C. Lee Giles, Yang Sun and Issac G. Council, “Measuring the Web Crawler Ethics,” WWW2010, ACM, 2010, pp. 1101-1102.

[2] Bhagwani J. and K. Hande, “Context Disambiguation in Web Search Results Using Clustering Algorithm”, International Journal of Computer Science and Communication, vol. 2, pp. 119-123.

[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

[5] Yang Sun,Ziming Zhuang and C. Lee Giles,” A Large- Scale Study of Robots.txt”, WWW2007, ACM, 2007, pp.1123–1124.

[6] 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.

[7] 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.

[8] Sullivan D., “Webspin: Newsletter “ http://contentmarketingpedia.com/Marketing-Library/Search/industryNewsSeptA1.pdf

[9] Vaughan L. and Thelwal M., “Search Engine Coverage Bias: Evidence and Possible causes”, Information Processing and Management, Vol 40, pp. 693-707.

[10] 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.

[11] C. Lee Giles, Yang Sun and Issac G. Council, “Measuring the Web Crawler Ethics,” WWW2010, ACM, 2010, pp. 1101-1102.

[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

[16] http://www.webmasterworld.com/search_engine_spiders/4348357.htm

[17] http://user-agent-string.info/list-of-ua/bot-detail?bot=bingbot

[18] http://whatis.riskyinternet.com/what-is/web-robot/discoveryengine-robot-6142/

[19] http://www.rhyolite.com/anti-spam/badbots.html

[20] http://support.google.com/webmasters/bin/answer.py?hl=en&answer=178852

[21] http://support.google.com/webmasters/bin/answer.py?hl=en&answer=182072

[22] http://www.majestic12.co.uk/projects/dsearch/

[23] http://www.bing.com/blogs/site_blogs/b/webmaster/archive/2009/08/10/crawl-delay-and-the-bing-crawler-msnbot.aspx

[24] http://help.yahoo.com/help/us/ysearch/slurp

[25] http://blocklistpro.com/content-scrapers/ahrefsbot-seo-spybots.html

[26] Kruskal,W. H., Wallis, W. A.”Use of Ranks in one-criterion Variance analysis”, Journal of the American Statistical Association, 47(260), 1952, pp.583-621.

[27] Paneerselvam, R.: Research Methodology. New Delhi: Prentice Hall of India Private Limited,2005.

[28] Ortega, J., L. And Aguillo, I,” Differences between web sessions according to the origin of their visits”,.Journal of Informetrics, 4, 2010,pp. 331-337 .