Mining Web Logs to Identify Search Engine Behaviour at Websites

Jeeva Jose
Department of Computer Applications, BPC College,
Mulakkulam North P.O, Piravom- 686664, Ernakulam District, Kerala, India,
E-mail: vijojeeva@yahoo.co.in

P. Sojan Lal
School of Computer Sciences, Mahatma Gandhi University,
Kottayam, Kerala, India
E-mail: padikkakudy@gmail.com

Keywords: web logs, web usage mining, search engines, crawler

Received: April 3, 2013

Web Usage Mining also known as Web Log Mining is the extraction of user behaviour from web log data. The log files also provide immense information about the search engine traffic at a website. This search engine traffic is helpful to analyse the ethics of search engines, quality of the crawlers, periodicity of the visits and also the server load. Search engine crawlers are automated programs which periodically visit a website to update information. Crawlers are the main components of a search engine and without them the websites will not be listed in the search results. The visibility of the web sites depends on the quality of the crawlers. Different search engines may have different behaviour at web sites. We intend to see the differences in behaviour of search engines in terms of the number of visits and the number of pages crawled. The hypothesis was tested and it was found that there is a significant difference in the behaviour of search engines.

Povzetek: Analizirano je obnašanje različnih spletnih iskalnih algoritmov.

1 Introduction

Web Usage Mining is the extraction of information from web log files generated when a user visits the website [1]. Web mining tasks include mining web search engine data, analysing web's link structures, classifying web documents automatically, mining web page semantic structures and page contents, mining web dynamics (mining log files), building a multilayered and multidimensional web. Web log data is usually mined to study the user behaviour at websites. It also contains immense information about the search engine traffic. The user traffic is removed by pre processing tasks, otherwise it may bias the search engine behaviour. The crawler is an important module of a web search engine. The quality of a crawler directly affects the searching quality of web search engines.

The process of identifying the web crawlers is important because they can generate 90% of the traffic on websites [2]. Commercial search engines play a vital role in accessing web sites and wider information dissemination [3, 4]. Search engines use automated programs called web crawlers to collect information from the web. These web crawlers are also known as spiders, bots, robots etc. These crawlers are highly automated and seldom regulated manually [5, 6, 7]. The crawlers periodically visit the websites to update the content. Certain web sites like stock market sites or online news may need frequent crawling to update the search engine repositories. Web crawlers access the websites for diverse purpose which includes security violations also. Hence they may lead to ethical issues like privacy, security and blocking of server access. Crawling activities are regulated from server side with the help of Robots Exclusion Protocol. This protocol is present in a file called robots.txt. Usually ethical crawlers first access this file which will be present at the root directory of the website and follow the rules specified by robots.txt [8, 9]. But it is also possible to crawl the pages at a website without accessing the robots.txt. Certain crawlers seems to disobey the rules in robots.txt after its modification because crawlers like "Googlebot", "Yahoo! Slurp" , "MSNbot" cache the robots.txt file for a website [8]. The web site monitoring software Google Analytics does not track crawlers or bots. This is because Google Analytics tracking is activated by a JavaScript that is placed on every page of the website. A crawler hardly recognizes these scripts and hence the visits from search engines are not recognized. In this work we intend to see whether all the search engines are behaving in the same way when it accesses a website.

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) user's IP address b) user's 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 [11, 12]. 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. Like the user access the website using a browser, the search engines also deploy user agents to access the web.

2 Background literature

Most of the works in Web Usage Mining is related to user behaviour. This is because websites like e-commerce websites will be interested in studying user behaviour for marketing, online sales and personalization. Several data mining tasks like clustering, classification, association rule mining etc. has been done for web log data of user behaviour. The web crawler ethics are measured to discover the ethicality of commercial search engine crawlers [9]. A survey of the use of the Robots Exclusion Protocol on the web through statistical analysis of a large sample of robots.txt files is done [10]. An empirical pilot study on the relationship between JavaScript usage and web site visibility was carried out to identify whether JavaScript based hyperlinks attract or repel crawlers resulting in an increase or decrease in web site visibility [6]. Another study is done with commercial search engines to find whether there is a significant difference in their coverage of commercial web sites [4]. A report on search engine ratings in United States is also available [3].

2.1 Preprocessing

The two data sets were extracted and it was found that the dataset 1 consists of 5,29,175 records for 8 weeks and dataset 2 consists of 2,60,775 records. The entries with unsuccessful status code 400 were eliminated. The HTTP requests with POST and HEAD was also removed. In addition all the user requests were removed to get the search engine requests. This is required as a user request in the input file may bias the results of search engine behaviour. After pre processing the resultant file contained only the successful search engine requests. Various search engine crawlers were identified. Some crawlers were identified from the IP address field. It contained substrings like “googlebot”, “baiduspider”, “msnbot” etc. The user agents were also helpful in identifying the bots or crawlers like Ezooms, discobot etc. Certain search engine crawlers with number of visits less than 5 per week was removed as it was considered irrelevant.The bots Ahrefbot, Seexie.com_bot, Turnitinbot, Yrspider were some of the bots in data set 1 whose number of visits were less than 5 in a week. For data set 2 the Alexabot was considered irrelevant. The crawlers in dataset 1 like Baiduspider, Discobot, Exabot, Feedtetcher-Google, Feedseeker, Gosospider, Ichiro, Magpie, MJ12bot, MSNbot, Seexie.com_bot, Slurp, Sogou, Sosospider, SpBot, Turnitinbot, Yahoo, Yeti, Yodao, Youdao and YrSpider were not present in dataset 2. After pre processing there were 22 crawlers for data set 1 and 5 crawlers for data set 2. The results for the number of visits made by various search engines of data set 1 is given in Table 1 and for data set 2 is given in Table 2.

We also intend to see the number of pages crawled by various search engines to see the dynamic behaviour of different search engines. Most of the search engines initially accessed the robots.txt file before crawling other pages except a few. Certain search engines crawled more pages compared with other bots or crawlers. For example the crawlers like Googlebot, Slurp, Bingbot, Feedfetcher-google, MJ12 etc crawled more number of pages and showed consistency in their behaviour. Table 3 shows the number of pages crawled by various search engines for data set 1 and Table 4 shows the result for data set 2.

2.2 Kruskal Wallis H test

Kruskal Wallis H Test detects if n data groups belong or not to the same population [13, 14]. 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 [15]. The formula for H static of Kruskal- Wallis test is given below where K is the number of samples.

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

where Rj is the sum of the ranks of the sample j, n_j 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.

Case I

H_0: There is no significant difference between the number of visits made by various search engine crawlers.
H_1: There is significant difference between the number of visits made by various search engine crawlers.

From the test statistic in Table 5, both the data sets show a clear evidence of rejecting the null hypothesis. 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 number of visits made by various search engines.

Case II

H_0: There is no significant difference between the number of pages crawled by various search engine crawlers.
H_1: There is significant difference between the number of pages crawled by various search engine crawlers.

Table 1: No: of visits by various crawlers for data set 1.

		Week										
No	**Crawler**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**μ**	**σ**
1	Alexa	1	5	10	1	2	0	2	3	24	3.00	3.207
2	Baiduspider	128	222	65	89	124	67	66	47	808	101.00	56.87
3	Bingbot	157	166	159	175	126	100	118	96	1097	137.13	30.94
4	Discobot	113	33	0	21	24	52	5	69	317	39.63	37.42
5	Exabot	1	1	2	1	5	3	3	3	19	2.38	1.408
6	Ezooms	50	48	40	22	0	23	38	41	262	32.75	16.74
7	Feedfetcher-Google	179	170	167	223	192	191	187	188	1497	187.13	17.28
8	Googlebot	211	226	238	273	212	207	200	207	1774	221.75	23.99
9	Gosospider	26	10	1	0	0	0	0	0	37	4.63	9.303
10	Ichiro	117	81	122	146	0	42	21	33	562	70.25	53.8
11	Magpie	20	17	13	15	13	15	14	18	125	15.63	2.504
12	MJ12bot	38	36	37	50	37	37	37	41	313	39.13	4.643
13	MSNbot	24	17	11	19	15	12	18	15	131	16.38	4.138
14	Slurp	149	114	144	190	144	145	160	145	1191	148.88	21.07
15	Sogou	48	34	37	54	40	44	43	60	360	45.00	8.701
16	Sosospider	28	31	42	38	31	32	30	28	260	32.50	4.957
17	SpBot	3	3	3	4	2	2	1	1	19	2.38	1.061
18	Yandex	51	71	57	72	102	44	51	74	522	65.25	18.64
19	Yahoo	22	0	0	0	0	1	1	0	24	3.00	7.69
20	Yeti	3	4	1	4	3	2	4	4	25	3.13	1.126
21	Yodao	16	59	26	100	72	42	10	32	357	44.63	30.6
22	Youdao	2	4	1	1	18	1	3	0	30	3.75	5.898

Table 2: No: of visits by various crawlers for data set 2.

		Week										
No	**Crawlers**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**μ**	**σ**
1	Ahrefsbot	79	0	1	19	37	66	31	48	281	35.13	28.6
2	Bingbot	31	41	27	43	23	30	28	17	240	30	8.64
3	Ezooms	3	20	26	38	26	24	9	28	174	21.75	11.1
4	Googlebot	42	49	42	44	42	49	35	60	363	45.38	7.41
5	Yandex	35	10	67	88	6	7	3	12	228	28.5	32.3

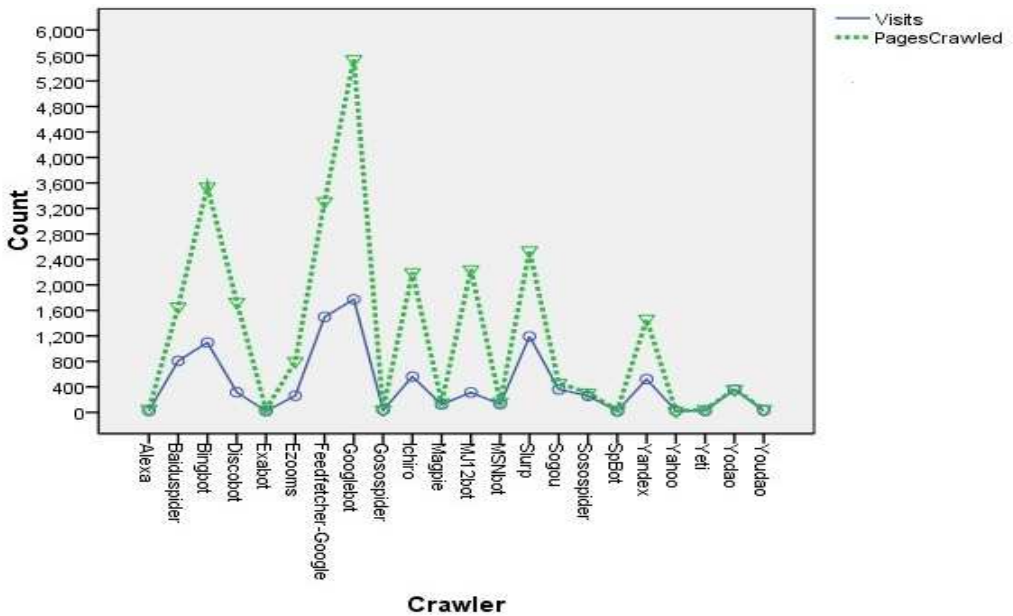

Figure 1: Time series sequence plot for data set 1.

Table 3. No: of pages crawled by various crawlers for data set 1

		Week										
No	**Crawler**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**μ**	**σ**
1	Alexa	2	13	27	2	4	0	4	4	56	7.00	8.96
2	Baiduspider	219	674	102	124	260	98	94	90	1661	207.63	199.03
3	Bingbot	368	559	519	526	404	232	287	647	3542	442.75	143.30
4	Discobot	889	161	0	119	92	289	6	178	1734	216.75	287.42
5	Exabot	2	11	4	2	11	6	5	6	47	5.88	3.52
6	Ezooms	235	160	77	57	65	59	83	67	803	100.38	63.79
7	Feedfetcher-Google	386	343	340	493	442	447	443	417	3311	413.88	53.81
8	Googlebot	841	895	682	847	655	525	540	556	5541	692.63	150.42
9	Gosospider	34	11	1	0	0	0	0	0	46	5.75	12.03
10	Ichiro	230	277	387	414	320	234	45	291	2198	274.75	113.86
11	Magpie	23	21	18	23	16	16	18	22	157	19.63	2.97
12	MJ12bot	174	304	224	392	255	285	294	316	2244	280.50	65.06
13	MSNbot	31	24	13	28	17	15	18	18	164	20.50	6.44
14	Slurp	367	253	297	410	310	264	308	331	2540	317.50	51.79
15	Sogou	72	42	47	61	52	54	51	80	459	57.38	12.89
16	Sosospider	32	38	57	42	36	36	35	33	309	38.63	8.03
17	SpBot	6	6	6	8	4	4	2	2	38	4.75	2.12
18	Yandex	140	250	99	171	216	102	212	276	1466	183.25	66.20
19	Yahoo	22	0	0	0	0	0	0	0	22	2.75	7.78
20	Yeti	6	9	2	7	7	4	7	7	49	6.13	2.17
21	Yodao	16	59	27	102	75	43	10	34	366	45.75	31.29
22	Youdao	4	8	2	2	25	2	7	2	52	6.50	7.86

Table 4: No: of pages crawled by various crawlers for data set 2.

		Week										
No	**Crawler**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**μ**	**σ**
1	Ahrefsbot	282	0	1	19	108	119	46	74	649	81.13	93.08
2	Bingbot	66	172	158	251	102	90	78	48	965	120.63	68.03
3	Ezooms	3	23	35	51	32	36	9	40	229	28.63	16.08
4	Googlebot	74	92	83	99	90	95	65	83	681	85.13	11.33
5	Yandex	39	18	123	199	6	7	4	13	409	51.13	71.65

The test statistic in Table 6 also shows that there is significant difference in the number of pages crawled by various search engines. The p-value for both the datasets is a strong evidence of rejecting the null hypothesis. A time series sequence plot was done for both data sets with total number of visits and total number of pages crawled. The result for data set 1 is shown in Figure 1 and for data set 2 is shown in Figure 2. We also intend to see whether there exists any correlation between the number of visits and number of pages crawled. The Karl Pearson's Correlation Coefficient [14] was calculated for both data sets. The data set 1 showed a strong positive correlation of 0.932 whereas the data set 2 showed a moderate positive correlation of 0.505.

3 Conclusion

The obtained results point to the differences in the behaviour of web crawlers by various search engines. The more the number of search engines accessing a website, the more will be its visibility when searching for a particular web site. The observed results show that all search engine crawlers are not visiting all the websites. In our experiment the data set 1 was accessed by more number of search engines compared to data set 2. Certain search engines were consistent in the number of visits and number of pages crawled while a few were not consistent or irregular in their visits and pages crawled. It is found that data set 1 is more visible to search engine crawlers as it is crawled by more number of search engines compared to data set 2. The results also showed a positive correlation between the number of visits and number of pages crawled. A better search engine optimization policy can be followed to make the websites visible to different search engines so that the websites will be listed top in the search engine rankings.

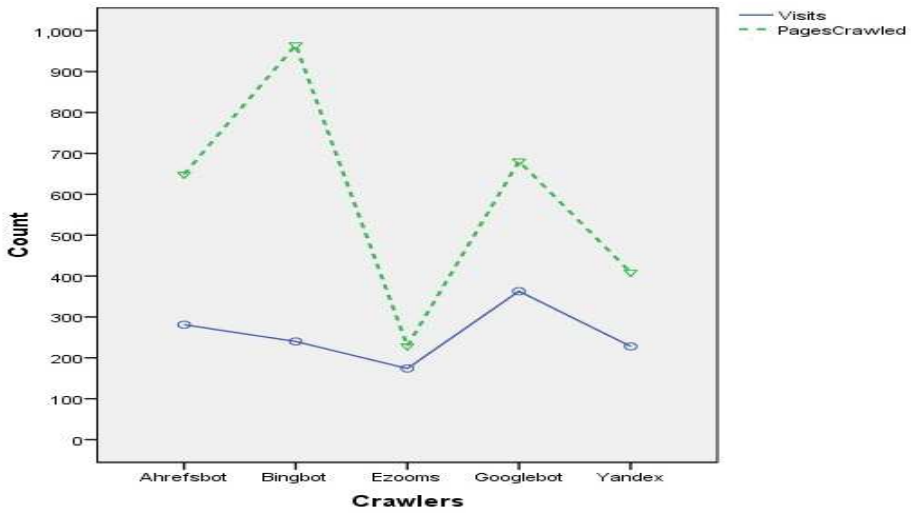

Figure 2: Time series sequence plot for data set 2.

Table 5: Test Statistic for Case I.

Kruskall Wallis Test		
	Data Set 1	**Data Set 2**
α	0.01	0.01
p-value	0.0001	0.044
Chi-square	148.734	9.799
df	21	4

Table 6: Test Statistic for Case II.

Kruskall Wallis Test		
	Data Set 1	**Data Set 2**
α	0.01	0.01
p-value	0.0001	0.013
Chi-square	154.85	12.714
df	21	4

Acknowledgement

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

References

[1] Kosala, R. And Blockeel, H., Web Mining Research: A Survey. ACM SIGKDD Explorations. 2(1), pp. 1-15, 2000.

[2] Mican, D. And Sitar-Taut, D., Preprocessing and Content/Navigational Pages Identification as Premises for an Extended Web Usage Mining Model Development. *Informatica Economica*, 13(4), pp. 168-179, 2009.

[3] Sullivan, D.2003, Webspin : Newsletter [online]. Available from: http://contentmarketingpedia.com/Marketing-Library/Search/industryNewsSeptA1.pdf .Accessed December 4, 2012.

[4] Vaughan, L. And Thelwall, M., Search Engine Coverage Bias: Evidence and Possible Causes, *Information Processing and Management*, 40(4), pp. 693-707, 2004.

[5] Bhagwani, J. And Hande, K., Context Disambiguation in Web Search Results Using Clustering Algorithm. International Journal of Computer Science and Communication, 2(1), pp. 119-123, 2011.

[6] Schwenke, F. And Weideman, M., The influence that JavaScript has on the visibility of a website to search engines – a pilot study. Informatics & Design Papers and Reports, 11(4), pp. 1-10, 2006.

[7] Thelwall, M., A Web Crawler Design for Data Mining, Journal of Information Science, 27(5), pp. 319-325, 2001.

[8] Drott, M, Indexing aids at corporate websites: The use of robots.txt and meta tags. Information Processing and Management, 38(2), pp. 209-219, 2002.

[9] Lee Giles, C., Sun, Y and Council, G., I., Measuring the Web Crawler Ethics. In: Proceedings of WWW 2010, ACM, pp. 1101-1102, 2010.

[10] Sun, Y. Zhuang, Z. .and Lee Giles, C., A Large-Scale Study of Robots.txt. In: Proceedings of WWW2007, ACM, pp. 1123-1124, 2007.

[11] Wahab, M.H,A, Mohd, M.N.H, Hanafi, H. F. Mohsin, M. F.M., Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm. In: Proceedings of World Academy of Science Engineering and Technology, pp.190-196, 2008.

[12] Spiliopoulou, M., Web Usage Mining for Web Site Evaluation. *Communications of the ACM*, 43(8), pp. 127-134, 2000.

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

[14] Paneerselvam, R., Research Methodology. New Delhi, Prentice Hall of India Private Limited, 2005.
[15] Ortega, J., L. And Aguillo, I., Differences between web sessions according to the origin of their visits, Journal of Infometrics, 4, pp. 331-337, 2010.