A Forecasting Model for the Pages Crawled by Search Engine Crawlers at a Web Site

Jeeva Jose
Department of Computer Applications
BPC College, Piravom
Kerala, India

P. Sojan Lal
School of Computer Sciences
Mahatma Gandhi University
Kottayam, Kerala, India

ABSTRACT

World Wide Web is exploding in terms of the number of web sites and users. Without search engines the web sites will not be visible to the users. Different search engine crawlers behave in different ways while they access a web site. The number of visits and pages crawled by search engines could be helpful in identifying their behavior and also the server load. A forecasting model in time series has been proposed for predicting the number of pages crawled by search engines. This model was compared with the actual values and it was found feasible.

General Terms

Web log mining, Web analytics

Keywords

Web sites, Web logs, Search engines, Crawlers

1. INTRODUCTION

Web mining tasks include mining web search engine data, analyzing 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 logs contain immense information about user behavior and search engine behavior at web sites. These logs are generated when a user visits a web site or search engine crawlers access a web site. Search engines employ web crawlers for collecting information from the web. These crawlers are highly automated and never regulated manually [1][5]. Crawlers are also known as spiders, bots, robots etc. These crawlers periodically visit World Wide Web and collect information which is spread across various pages. Certain pages like online news pages, stock market pages may be updated frequently whereas other pages may be seldom modified. 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 working of a search engine crawler is explained in [2]. Search engine crawlers should be flexible with low cost and high performance. A good crawler should be able to deal with the odd behavior of servers, system crashes and bad HTML pages. Certain crawlers are ethical in their behavior while many are not. Because of the highly automated nature of the robots, rules must be made to regulate such crawling activities to manage the server workload and denying access to confidential or private information [3]. A file called robots.txt is placed at the root of the web site directory which specifies the Robots Exclusion Protocol. These are a set of rules that the robots are expected to follow. The ethical robots initially read the robots.txt file and follow the regulations while non ethical crawlers access the web pages without reading the instructions in this file. 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 [4]. A forecasting model in time series was constructed to predict the number of pages crawled by various search engines.

2. RELATED WORK

There are several works in web usage mining which specifies the behavior of users. Most of the works concentrate on user behavior since it has application in many e-commerce web sites. There are open source software available like Google Analytics which measure 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. A large scale study of robots.txt is done by Sun et al [3] and the ethics of web crawlers is studied by Giles [1]. Schwenke et al has performed a study on the relationship between JavaScript usage and web site visibility to identify whether JavaScript based hyperlinks attract or repel crawlers resulting in an increase or decrease in web site visibility is done by [6].Another study is performed with commercial search engines to find whether there is a significant difference in their coverage of commercial web sites [7]. A study report on search engine ratings in United States is also available [8]. An investigation on the revisitation patterns in World Wide Web navigation is done but it does not deal with the search engine crawler's revisits [9]. Our intention is to propose a model for the prediction of the behavior of search engine crawlers in terms of the number of pages crawled.

3. METHODOLOGY

3.1 Format of Log Files

Web logs are maintained by web servers and contain information about search engine crawlers and users accessing the site. Logs are mostly stored as simple text files, each line corresponding to one access. 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[10][11]. 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 [10].

3.2 Pre Processing

The first pre processing task is data cleaning. The process of data cleaning is to remove noise or irrelevant data. Web server access logs represent the raw data source. It is important to identify and discard the data recorded by user visits, images, sounds, java scripts etc that is often redundant and irrelevant [12]. We need to extract only the web robots or web crawlers and remove user visits otherwise it may bias the behavior of search engine crawlers. 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.

Search engine crawlers can be identified from their IP address, user agent used for crawling. The total number of pages crawled by various search engine crawlers for each day was obtained. 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 45 days ranging from April 1, 2011 to May 15, 2011 and second dataset belongs to an academic website www.bpccollege.ac.in ranging from November 1, 2012 to December 15, 2012 comprising of 45 days. After extraction, there were 4,28,345 records for data set 1 and 2,38,446 records for data set 2. The successful search engine requests were filtered for further processing.

3.3 Forecasting Model

A forecast is an estimate of an event which will happen in future. The event may be the demand of a product or growth of a technology. The forecast value is not a deterministic quantity and only an estimate based on the past data related to a particular event. There are several forecasting techniques available. The shape of the curve representing the forecast of an event is one or more of the patterns [13].

a) Horizontal pattern
b) Trend pattern
c) Seasonal pattern
d) Cyclical pattern
e) Random pattern

a) Horizontal pattern

A horizontal pattern exists when the data values fluctuate around a constant mean. Figure 1 shows a horizontal pattern.

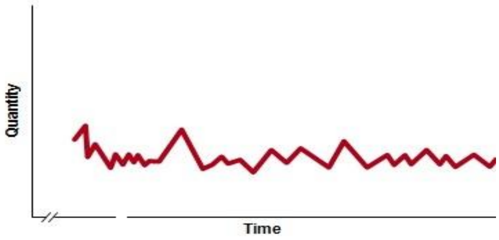

Figure 1. Horizontal pattern

b) Trend pattern

In a trend pattern, data is progressively increasing or decreasing with time. Figure 2 shows a trend pattern

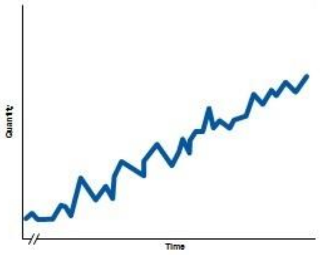

Figure 2. Trend Pattern

c) Seasonal pattern

The seasonal demand exists when the demand fluctuates according to some seasonal factors. Data exhibits a regularly repeating pattern at constant intervals. Figure 3 shows the seasonal pattern.

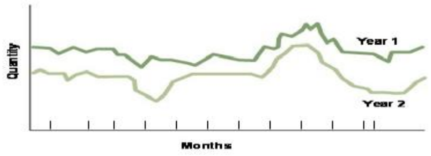

Figure 3. Seasonal Pattern

d) Cyclical pattern

The cyclical pattern shows the increase or decrease of data with time. Figure 4 shows cyclical pattern.

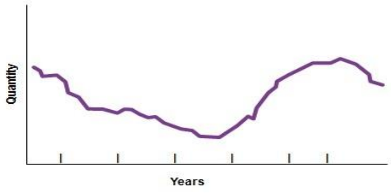

Figure 4. Cyclic pattern

e) Random pattern

Random pattern may happen due to random process. This may exist in some situations for which no reason can be given.

The web log data extracted from both data sets showed that the patterns for the number of pages crawled by various search engines exhibited random pattern. The data was also noisy and hence normalization was done before forecasting. There are several methods for normalization and the data was normalized using Min-Max normalization [15]. It is calculated as follows.

$$\grave{a} = \frac{a - \mathrm{minA}}{\mathrm{maxA} - \mathrm{minA}}\left(\mathrm{new}_{\mathrm{maxA}} - \mathrm{new}_{\mathrm{minA}}\right) + \mathrm{new}_{\mathrm{minA}} \qquad (1)$$

where minA = Minimum value from the observed values
maxA = Maximum value from the observed values
newmaxA = maximum value for the new range
newminA = minimum value for the new range
a= value to be mapped
à = normalized value

The data values are mapped to a range [0, 1]. Single exponential smoothing method is a forecasting method used when the data shows a random pattern [14]. There are double exponential and triple exponential smoothing methods available. Double exponential smoothing method is used whenever there is a trend in the data. Triple exponential smoothing method when there is seasonal changes as well as trend. Since both our data sets do not showed any trend, single exponential smoothing method was chosen. This is a widely used method and requires very little data. The forecast of the period t is computed by applying some correction over the forecast value of the immediate preceding period. The correction quantity is a portion α of the difference between the actual value of the immediate preceding period and the forecast of the immediate preceding period. The exponential smoothed forecast is computed as given below.

$$Ft = Ft\text{-}1 + \alpha(Dt\text{-}1 - Ft\text{-}1) \qquad (2)$$

where Ft is the smoothed average forecast of the period t, Ft-1 is the smoothed average forecast of the period (t-1), Dt-1 is the actual value of the period (t-1) and α is the smoothing constant (0< α<1). Larger the α gives the more responsive the forecast. We have taken α = 0.8. For initial seed we have taken the moving average of first two values. The results of the preprocessing and exponential smoothed forecasts for both data sets 1 and 2 is given in Table I. The Forecast Error (FE) is calculated as

$$FE = \grave{a} - Ft \qquad (3)$$

The Squared Forecast Error (SE) is calculated as

$$SE = (\grave{a} - Ft)2 \qquad (4)$$

The Mean Forecast Error (MFE) is computed as

$$MFE = \sum_{t=1}^{N} \frac{(\grave{a} - \mathrm{Ft})}{N} \qquad (5)$$

The MFE for data set 1 is 0.02 and for data set 2 is 0.01. The Mean Squared Error (MSE) is calculated as follows.

$$MSE = \sum_{t=1}^{N} \frac{(\grave{a} - \mathrm{Ft})2}{N} \qquad (6)$$

The Mean Squared Error (MSE) for data set 1 is 0.04 and for data set 2 is 0.12. The actual forecasted value is calculated as

$$\mathrm{Ft}' = \frac{\grave{a} * (\mathrm{maxA} - \mathrm{minA})}{\left(\mathrm{new}_{\mathrm{maxA}} - \mathrm{new}_{\mathrm{minA}}\right)} - \mathrm{new}_{\mathrm{minA}} + \mathrm{minA} \qquad (7)$$

Figure 5 and Figure 6 show the graphical representation of observed values and the forecasted values for data set 1 and data set 2 respectively.

Table1. Results of Pre processing and Exponential Smoothed Forecasts

Data Set 1							Data Set 2						
Day (X)	**Pages Crawled (Y)**	**à**	**F_t**	**FE**	**SE**	**F_t'**	**Day (X)**	**Pages Crawled (Y)**	**à**	**F_t**	**FE**	**SE**	**F_t'**
1	611	0.62					1	101	0.58				
2	722	0.81	0.71			664	2	145	0.86	0.72	-0.14	0.02	123
3	457	0.35	0.79	0.44	0.20	710	3	73	0.40	0.83	0.43	0.19	141
4	510	0.44	0.44	0.00	0.00	508	4	39	0.18	0.49	0.30	0.09	87
5	677	0.73	0.44	-0.29	0.09	510	5	33	0.15	0.25	0.10	0.01	49
6	598	0.59	0.67	0.08	0.01	644	6	88	0.50	0.17	-0.33	0.11	36
7	613	0.62	0.61	-0.01	0.00	607	7	39	0.18	0.43	0.25	0.06	78
8	794	0.94	0.62	-0.32	0.10	612	8	89	0.50	0.23	-0.27	0.07	47
9	657	0.70	0.87	0.18	0.03	758	9	16	0.04	0.45	0.41	0.17	81
10	452	0.34	0.73	0.39	0.15	677	10	62	0.33	0.12	-0.21	0.04	29
11	627	0.65	0.42	-0.23	0.05	497	11	16	0.04	0.29	0.25	0.06	55
12	830	1.00	0.60	-0.40	0.16	601	12	16	0.04	0.09	0.05	0.00	24
13	634	0.66	0.92	0.26	0.07	784	13	87	0.49	0.05	-0.44	0.20	18
14	718	0.80	0.71	-0.09	0.01	664	14	63	0.34	0.40	0.06	0.00	73
15	402	0.25	0.79	0.53	0.28	707	15	24	0.09	0.35	0.26	0.07	65
16	533	0.48	0.36	-0.12	0.01	463	16	79	0.44	0.14	-0.30	0.09	32
17	427	0.30	0.46	0.16	0.03	519	17	104	0.60	0.38	-0.22	0.05	70
18	402	0.25	0.33	0.08	0.01	445	18	85	0.48	0.55	0.08	0.01	97
19	436	0.31	0.27	-0.04	0.00	411	19	29	0.12	0.49	0.37	0.14	87
20	462	0.36	0.30	-0.05	0.00	431	20	125	0.73	0.20	-0.54	0.29	41
21	427	0.30	0.35	0.05	0.00	456	21	53	0.27	0.63	0.35	0.12	108
22	294	0.06	0.31	0.24	0.06	433	22	167	1.00	0.34	-0.66	0.43	64
23	403	0.25	0.11	-0.14	0.02	322	23	98	0.56	0.87	0.31	0.10	146
24	394	0.24	0.23	-0.01	0.00	387	24	117	0.68	0.62	-0.06	0.00	108
25	425	0.29	0.24	-0.06	0.00	393	25	85	0.48	0.67	0.19	0.04	115
26	577	0.56	0.28	-0.28	0.08	419	26	143	0.85	0.52	-0.33	0.11	91
27	489	0.40	0.50	0.10	0.01	545	27	152	0.90	0.78	-0.12	0.02	133
28	557	0.52	0.42	-0.10	0.01	500	28	53	0.27	0.88	0.61	0.37	148
29	462	0.36	0.50	0.15	0.02	546	29	10	0.00	0.40	0.40	0.16	72
30	303	0.08	0.39	0.31	0.09	479	30	28	0.11	0.08	-0.04	0.00	22
31	314	0.10	0.14	0.04	0.00	338	31	64	0.34	0.11	-0.24	0.06	27
32	386	0.22	0.11	-0.12	0.01	319	32	48	0.24	0.30	0.05	0.00	57
33	626	0.64	0.20	-0.44	0.20	373	33	10	0.00	0.25	0.25	0.06	50
34	471	0.37	0.55	0.18	0.03	575	34	122	0.71	0.05	-0.66	0.44	18
35	565	0.54	0.41	-0.13	0.02	492	35	22	0.08	0.58	0.50	0.25	101
36	446	0.33	0.51	0.18	0.03	550	36	137	0.81	0.18	-0.63	0.40	38
37	410	0.27	0.37	0.10	0.01	467	37	43	0.21	0.68	0.47	0.22	117
38	421	0.28	0.29	0.00	0.00	421	38	47	0.24	0.30	0.07	0.00	58
39	402	0.25	0.29	0.03	0.00	421	39	87	0.49	0.25	-0.24	0.06	49
40	388	0.23	0.26	0.03	0.00	406	40	14	0.03	0.44	0.42	0.17	79
41	372	0.20	0.23	0.03	0.00	392	41	121	0.71	0.11	-0.60	0.36	27
42	486	0.40	0.21	-0.19	0.04	376	42	45	0.22	0.59	0.36	0.13	102
43	446	0.33	0.36	0.03	0.00	464	43	28	0.11	0.30	0.18	0.03	56
44	564	0.53	0.33	-0.20	0.04	450	44	26	0.10	0.15	0.05	0.00	34
45	327	0.12	0.49	0.37	0.14	541	45	105	0.61	0.11	-0.49	0.24	28

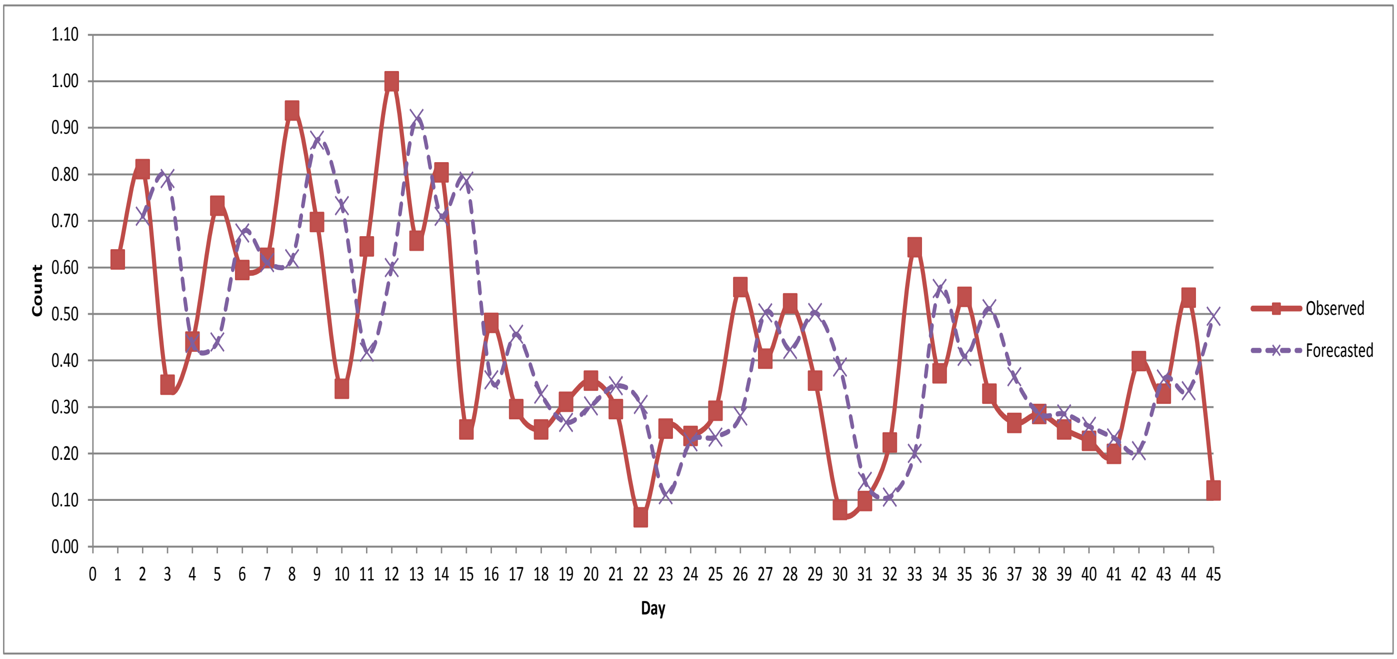

Figure 5. Observed and Forecasted values for the number of pages crawled at web site 1

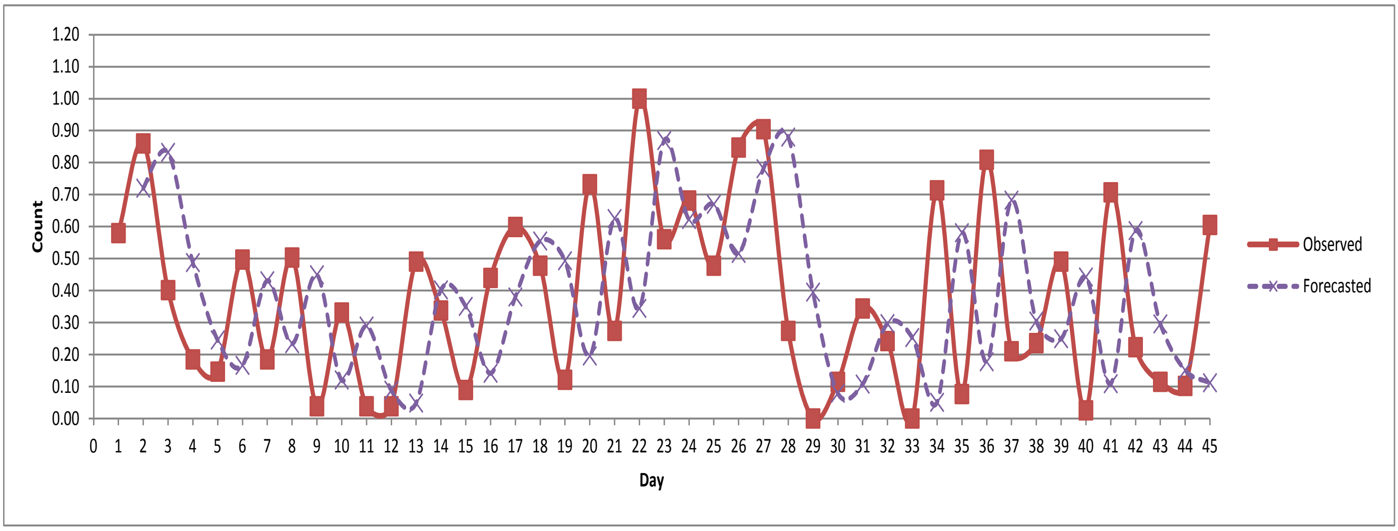

Figure 6. Observed and Forecasted values for the number of pages crawled at web site 2

4. CONCLUSION

The results showed that for both the data sets, the single exponential smoothing method works well for forecasting the number of pages crawled at web sites. This forecasting model can be used in analyzing and predicting the server load. Since the data is noisy and does not showed any of the patterns, this model is well suited for making predictions for the web log data like the number of pages crawled at a web site. The Mean Square Error (MSE) and Mean Forecast Error (MFE) was very less for both the data sets which is an indication of the acceptance of this model.

5. ACKNOWLEDGMENT

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.

6. REFERENCES

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

[2] 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.

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

[4] Animesh Tripathy, Prashanta K Patra, “A Web Mining Architectural Model of Distributed Crawler for Internet Searches Using PageRank Algorithm”, Proceedings of the Asia-Pacific Services Computing Conference, IEEE,2008.

[5] 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.

[6] 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.

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

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

[9] Linda T. and Saul Greenberg,”Revisitation Patterns in World Wide Web Navigation”, CHI, ACM, 1997, pp. 22-27.

[10] 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.

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

[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] Kothari C.R, Research Methodology Methods & Techniques, New Age International Publishers, Revised Second Edition, 2007.

[14] Pannerselvam R, Research Methodology, Prentice Hall of India Private Limited, 2005.

[15] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Elsevier, Third Edition,2012.