ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Analysis of Robot Detection Approaches for Ethical and Unethical Robots on Web Server Log

Mitali Srivastava*
Department of Computer Science, Institute of Science,
Banaras Hindu University,
Varanasi, India

Atul kumar Srivastava**
Department of Computer Science, Institute of Science,
Banaras Hindu University,
Varanasi, India

Rakhi Garg
Computer Science Section, Mahila Maha Vidayalya, Banaras Hindu University,
Varanasi, India

P. K. Mishra
Department of Computer Science, Institute of Science,
Banaras Hindu University,
Varanasi, India

*First author, **Corresponding Author

Abstract:Due to proliferation of Web robots, it is becoming important to detect robots on commercial and educational websites. Web robots not only consumes a significant proportion of network bandwidth but also a threat for privacy and security. There are various techniques to detect robot from server log *i.e*. checking robots.txt access, matching keywords in User agent field, calculating browsing speed etc. In this paper, we have implemented and compared theabove three robot detection approaches on real Web server log. Moreover, we have also applied combined approach to detect robots and found that combined approach works better than other three approaches.

Keywords: Robot detection, Web server log, Web usage mining, Business intelligence, Search engines, Data extraction

I. INTRODUCTION

Web robots, spiders, crawlers, indexer or Web wanderers, are software programs that traverses hyperlink structure of Web site to extract Web resources. Web robots are mainly used by search engines to retrieve Web page and index them to their databases [1]. Whenever a Web robots retrieve a resource from a Web site its entries are recorded into a Web server log file. There are basically two types of Web robots *i.e.* Ethical and Unethical. *Ethical robots* follow standard guidelines provided by robots designers while traversing a Web site. On other hand, *Unethical robots* do not follow standard guidelines and try to hide their entries in Web server log. [2].Apart from extracting Web resources**,** other types of functions are also performed by different types of robots *i.e*. *Line checkers* that are used by Web site administrator to detect broken links in the Web site; *Offline browsers* that are used by all common browsers to download some set of resources related to a Web site for future purpose; *Shop bots* and *Price bots* that are used to monitor and compare product prices on e-commercial web sites. In addition to above functions, Web robots can also be used for some malicious purposes like sending spam emails, stealing Business intelligence etc. [3, 4]. Sometimes robots are useful but there are situations where it is necessary to detect robots. These situations are:-

- It is being difficult to apply Web usage mining techniques on server log file due to presence of large number robots requests.
- Sometimes Web robots consume larger part of network bandwidth that slows down the speed of server response.
- Several e-commercial web sites also want to stop unauthorized access of stealing their business intelligence information by web robots [1, 5, 6, and 7].

Web server log is one of the valuable source of information that stores the activities of user. Ethical robots can easily be identified from Web server log by using robots.txt method [8]. In this paper we have discussed various techniques used for detection of ethical and unethical robots requests from server log file and done the comparative analysis of these techniques. Section II describes various robot detection techniques applied and their limitations. After that, section III includes implementation and analysis of results obtained. At last section IV concludes the paper.

II. ROBOT DETECTION TECHNIQUES

There are various techniques to identify ethical and unethical robots from server log file which are discussed as follows:

A. *By checking request of robots.txt file*

According to robot exclusion protocol, every robot should request robots.txt file first while navigating through a Web site. This file is used by Web site administrator to set permissions on few Web resources to restrict their access from Web robots.[9] An ethical robot follow this guideline hence they can be easily detected by checking host who have requested robots.txt file in Web server log [10]. After identifying robots host, entries corresponding to this host are considered as robots requests. An example of robots.txt request is given in Figure 1. This entry is taken from Banaras Hindu University Web server.

199.30.20.11 - - [24/Mar/2014:00:29:51 +0530] "GET /robots.txt HTTP/1.1" 404 287 "-" "msnbot-media/1.1 (+http://search.msn.com/msnbot.htm)"
Figure 1. An entry in Apache server log by an Ethical robot

B. By using keywords matching in User agent field

Another important guideline for robots is to declare themselves in User agent string of server log while requesting a Web resource in a Web site. User agent is an important field of server log file that contain information regarding client's operating system and browser. Ethical robots put their name in User agent string in place of browser name. In recent years, due to tremendous increase in different types of Web robots, it is not possible to collect list of user agents of all well-known robots. Since some robots combine their name with keywords like robot, bot, crawler, indexer or spider so robots request can be detected by User agents that contain these keywords. For example: Msnbot, Googlebotetc [3, 5].

C. By calculating browsing speed

The above discussed two approaches are useful to detect ethical robots. Some unethical robots neither check robots.txt file nor specify their name in User agent string. One important approach to detect these kind of robots is by calculating browsing speed. Generally Web robots traverse website in fast and exhaustive manner. To apply this method, initially user's activities are grouped according to same host name then browsing speed is calculated by formula given in *Eq.* 1. If browsing speed is less than threshold θ then users' activates belong to a web robot [11].

$$Browsing\ speed = \frac{Total\ time\ spent\ on\ pages}{Number\ of\ pages} \qquad (1)$$

D. Combined approach

To detect both type of robots *i.e.* Ethical and unethical, we have combined the above discussed three approaches. In this approach we firstly detect robots requests by "matching keywords in User agent". After that robots are detected by looking host who have accessed "robots.txt" file first. Further, we have applied third method based on browsing speed to detect the robot.

III. EXPERIMENT AND RESULT ANALYSIS

We have implemented the above discussed four robot detection approaches by collecting access log file from Banaras Hindu University Website. After that datasets for 6 hours, 12 hours and 18 hours are extracted from access logs by applying our proposed data extraction algorithm .This algorithm extracts data according to particular duration with removal of duplicates [12].Description of datasets extracted by Data extraction algorithm are given in Table 1.

Further,we have applied four robot detection techniques i.e. by checking host of robots.txt file, by keywords matching in User agent field, by calculating browsing speed, combined approach on data sets D1, D2 and D3. All techniques are implemented in JAVA (JDK 1.7) language and executed on system having UBUNTU 14.04 operating system, Intel core I5 processor (4 cores) and 4GB RAM. Browsing speed is set to 2.0 seconds per page in our case.

Table 1: Description of datasets

Data sets	Duration	Size (MB)	Total Number of duplicate Requests	Total number of requests without duplicates
D1	24/03/2014:00:00:0 to 24/03/2014:05:59:59	8.9	423	37902
D2	24/03/2014:00:00:0 to 26/03/2014:11:59:59	71	7550	309047
D3	24/03/2014:00:00:0 to 30/03/2014:17:59:59	550	19504	809994

From Figure 2, it can be observed that, the minimum number of robots requests is generated by robots.txt method and maximum numbers of robots requests are generated by browsing speed method if they are applied alone on datasets D1, D2 and D3.In combined approach, number of generated robots requests are larger than all the three techniques for all given datasets. Figure 3 shows the contribution of robots requests generated by combined approach in total number of request. The total number of robots requests by using this approach are 25.62%, 20.06 % and 15.68% of total requests for datasets D1, D2 and D3.

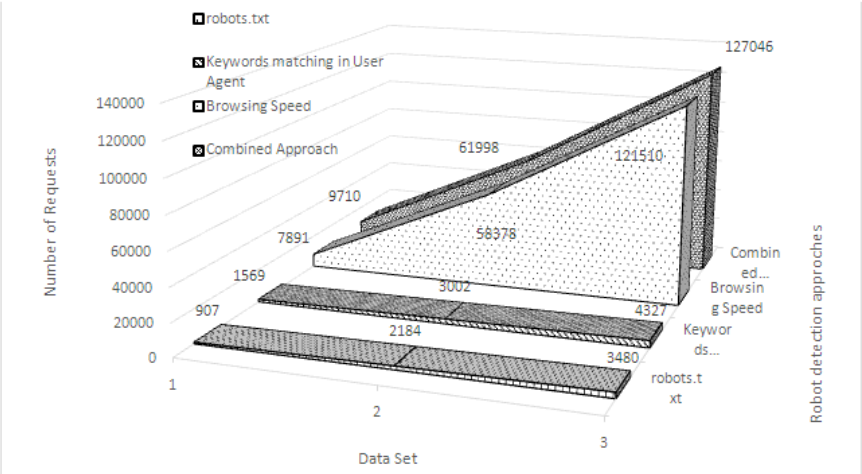

Figure 2. Number of robots request by various robot detection techniques

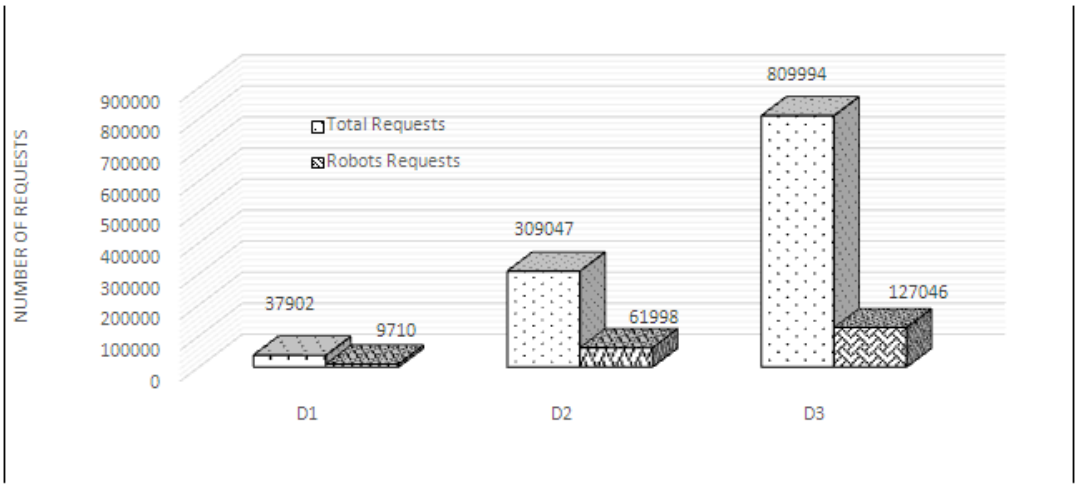

Figure 3: Proportion of robots requests by using combined technique

IV. CONCLUSION

There are various techniques to detect ethical and unethical robots requests from server log. In this paper we have compared four techniques *i.e*. robots.txt accesses, matching keywords in User agent string, browsing speed method and combined approach to detect robots requests. After performing an experiment on real Web server log, we have observed that, single technique alone is not sufficient to detect ethical and unethical robots. However, combined approach of above discussed three approaches works better and detects maximal number of robots requests than other three approaches.

REFERENCES

[1] Tan, Pang-Ning, and Vipin Kumar. "Discovery of web robot sessions based on their navigational patterns." Intelligent Technologies for Information Analysis. Springer Berlin Heidelberg, 2004. 193-222.

[2] P.-N. Tan, V. Kumar, Modeling of web robot navigational patterns, in: WEBKDD 2000—Web Mining for E-Commerce—Challenges and Opportunities, Second International Workshop, 2000.

[3] N. Geens, J. Huysmans and J. Vanthienen, "Evaluation of web robot discovery techniques: a benchmarking study," in Industrial Conference on Data Mining, 2006, pp. 121-130.

[4] Lu, Wei-Zhou, and Shun-zheng Yu. "Web robot detection based on hidden Markov model." 2006 International Conference on Communications, Circuits and Systems. 2006.

[5] Doran, Derek, and Swapna S. Gokhale. "Web robot detection techniques: overview and limitations." Data Mining and Knowledge Discovery 22.1-2 (2011): 183-210.

[6] Kwon, Shinil, Young-Gab Kim, and Sungdeok Cha. "Web robot detection based on pattern-matching technique." Journal of Information Science 38.2 (2012): 118-126.

[7] Srivastava, Mitali, Srivastava Atul Kumar, Rakhi Garg, and P. K. Mishra. “Comparative Analysis of Robots Detection Approaches.” International Journal of Advanced Research in Computer and Communication Engineering (2015).

[8] Srivastava, Mitali, Rakhi Garg, and P. K. Mishra. "Preprocessing techniques in web usage mining: A survey." International Journal of Computer Applications 97.18 (2014).

[9] Koster, M. (1994), “A standard for robot exclusion”, available at: www.robotstxt.org/orig.html

[10] D. Tanasa and B. Trousse. Advanced data preprocessing for intersites web usage mining. Intelligent Systems, IEEE, 19(2):59{65, 2004.

[11] G Castellano, A. Fanelli and M. Torsello, "LODAP: a log data preprocessor for mining web browsing patterns," in Proceedings of the 6th Conference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, 2007, pp. 12-17.

[12] Srivastava, Mitali, Rakhi Garg, and P. K. Mishra. "Analysis of Data Extraction and Data Cleaning in Web Usage Mining." Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015). ACM, 2015.