ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

Comparative Analysis of Robot Detection Techniques on Web Server Log

Mitali Srivastava[1], Atul Kumar Srivastava[2], Rakhi Garg[3], P. K. Mishra[4]

Department of Computer Science, Faculty of Science, Banaras Hindu University, Varanasi, India[1, 2, 4]

Computer Science Section, Mahila Maha Vidyalaya, Banaras Hindu University, Varanasi, India[3]

Abstract: Web robots are software programs which automatically traverse through hyperlink structure of Web to retrieve Web resources. Robots can be used for variety of tasks such as crawling and indexing information for search engines, offline browsing, shopping comparison and email collectors. Apart from that robots can also be used for some malicious purposes like sending spam mails, stealing business intelligence etc. It is necessary to detect robots due to privacy, security and performance of server related issues. Several well-known techniques to detect robots are : robots.txt check, known robot's IP address, User agent mapping, keywords matching in User agent field, browsing speed, unassigned referrer etc. In this paper we have discussed as well as implemented various robot identification techniques on real server log data and compared their performance for a given dataset.

Keywords: Robot detection, Web server log, Web usage mining, Data extraction.

I. INTRODUCTION

The exponential growth of Web based economy has led to explosive increase in volume of dynamic and time-sensitive information. Generally business organizations on Web consider this as a useful information in terms of user's perspective regarding products and services. This leads to a rapid increase of various types of robots on Web. Robot is a software program which automatically traverses hyperlink structure of Web site to retrieve information from Web [1].

Sometimes robots are called as spiders, bots, crawlers or Web wanders. Web robots are generally used for different purposes e.g. for resource discovery and indexing for search engines like Google, Yahoo etc.; as an offline browsers which downloads some set of resources for browsing; as a line checkers to check hyperlink validity; as a shopping comparison robots to monitor, compare specific product prices on other e-commercial Web site; and as an email collector to collect record of emails provided on web page [1, 3, 5]. Web robots can also be used by Web site administrator to solve maintenance issues like checking the broken hyperlinks and mirroring. However, some robots can also be programmed for malicious purposes like sending spam mails [4].

Following are the situation where it is required to identify the robots [1, 2, 3].

- Business organizations on Web wants to disable unauthorized access of robots to collect their business intelligence information.
- Web usage analysts/Researchers are willing to distinguish human user and robot to identify correct user's navigation behaviour.
- Sometimes Web robots consume larger part of network bandwidth that slows down the speed of server response.

Whenever a particular client i.e. human user or robot, request a particular resource on Web then its activity is automatically stored in a special file called server log file by Web server. This file is usually maintained by Web site administrator. Web robots can be identified by analyzing server log file [6].

This paper discusses various techniques used for robot detection from server log file and does their comparative analysis. Section II describes description of server log data and its attributes. Section III describes various robot detection techniques and their limitations. Further section IV includes implementation and analysis of results obtained. At last section V provides the conclusion of paper.

II. DATA COLLECTION AND DESCRIPTION OF ATTRIBUTES

Whenever a user traverse Web resources on a particular Web site, Web server stores all the information regarding user's activity into server log file. There are several types of server log which can be collected from server e.g. access log, referrer log, error log and agent log.

For analysis we have collected access logs of Banaras Hindu University Web server (Apache/2.2.3 Red Hat) for duration of one month. This Apache web server follows combined log format which is also referred as extended common log format (ECLF).

Example of one entry from access log is given in Fig. 1. The description of attributes is given in Table 1.

III. ROBOT DETECTION TECHNIQUES

There are several techniques to identify robots from server log file which is described as follows:-

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

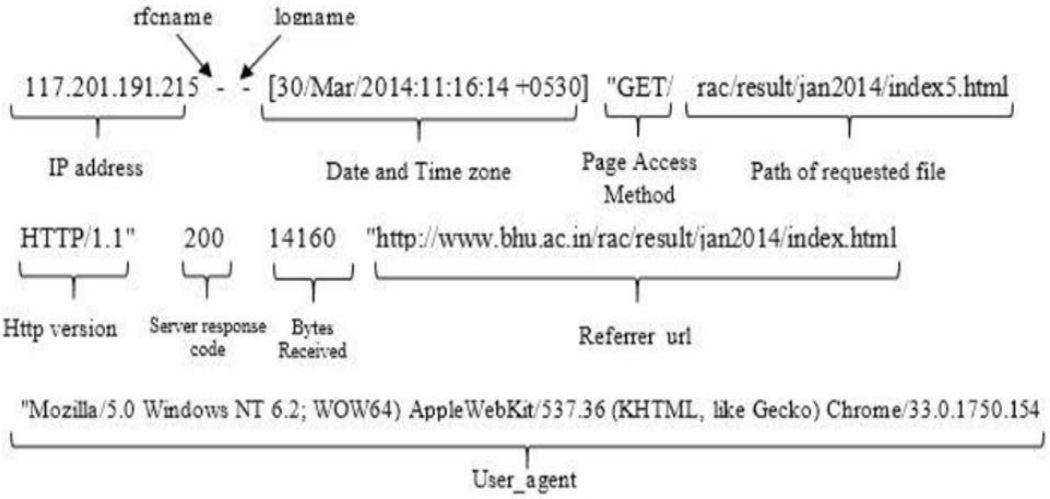

Figure 1: Sample log entry from Banaras Hindu University Web server log in ECLF format [6]

Table 1: Description of attributes of server log in ECLF format

Attributes	Description
IP address/host name	IP address/host name of client
rfcname	Authentication information of client
logname	Login name of client
Date and Time zone	Date and time of request with zone
Page Access Method/HTTP access method	Mode of HTTP request. The value can GET,POST,HEAD etc.
Path of requested file/Requested url	Path of requested resource on server
Http version	Version of HTTP protocol
Server response code/HTTP status code	Three digit code which depicts status of request
Bytes Received	Number of bytes transferred from the server
Referrer url	Name of URL from where request is done.
User_agent	Client's browser and operating system information
Cookies	Provides details of particular client if enabled by client and server both.

A. By checking request of robots.txt file

According to robot exclusion protocol by M. Koster, every robot which traverse for resources on Web should check robots.txt file first [8]. When a particular web site administrator wants to provide some instructions to robots they put a text file named as robots.txt in the root of Web site e.g. http://www.bhu.ac.in/robots.txt. This file actually sets the permissions for robots to access resources. The requests for robots.txt file can be used as an identifier for robot because human users are rarely interested in this file. However, since some robots does not follow robot exclusion protocol and traverse through pages without requesting robots.txt file so this method becomes unreliable in such situations [3]

B. By using known robot's IP addresses

Another method to detect robot's request is by checking IP address of request with list of known robot's IP addresses. Several web sites provide list of updated IP addresses of well-known robots e.g. Googlebot, YahooSlurp, MSNBot etc. Since many robots having dynamic IP addresses and routed over proxy server makes this technique unreliable. Apart from that this technique is effective only for popular robots which provides their updated list of IP [4].

C. By using User agent field mapping

Another guideline for robot designers is to declare themselves in User agent string. User agent string generally provides the information of client's operating system and browser version .Web robot should specify their names in user agent string in place of browser name unlike the user agent string of human web user but some unethical robots violets this guidelines by hiding their name in user agent string and filled it with standard browser name. Ethical robots can be identified by checking User agent field of server log with the list of user agent string of known robots. A list of known robot's user agent can also be find out from different web sites [4, 5].

D. By using keywords matching in User agent field

Previous technique is useful for some known robots whose User agents can be collected from websites. It is not possible to collect User agents for all robots because a single website cannot provide list of user agents for all robots. Since some robots combines their names with keywords "robots", "spider", "crawlers" and "bot" to create their User agent string so robots can also be

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

identified by matching these keywords in User agent field of log file [4].

IV. EXPERIMENT AND RESULT ANALYSIS

As it is discussed above access log files are collected from Banaras Hindu University website. Further, datasets are extracted from access logs for one day, three days and seven days by applying our proposed algorithm which can be referred from [7]. Extracted data sets are described in Table 2.

After that we have applied four robot detection techniques: - checking request of robots.txt file, known robot's IP addresses, User agent field mapping and keywords matching in User agent field on data sets. Further combination of these four techniques can also be used to detect robots. All techniques are implemented with JAVA (JDK 1.8) on system having UBUNTU 14.04 operating system, Intel core I3 processor and 6GB RAM.

List of IP addresses and User agents of known robots are collected from the website http://www.iplists.com. We have considered Google IP List, Yahoo IP List and Misc. IP List files for implementing techniques IP address and User agent mapping [9].

Fig. 2 shows comparative results of all four techniques. As it is clear from Fig. 2 that for all data sets larger number of robots are detected by using keywords matching in User agent field technique and smaller number of robots are detected by using robots.txt technique. Not all techniques alone sufficient to detect robot's requests so we have applied combination of all these four techniques.

Fig. 3 shows the result of the combination of these four techniques. From Fig. 2 and Fig.3 3 it is clear that result obtained from combined technique are slightly larger than the four technique i.e. *keywords matching in User agent field* on datasets.

If all robots follow robot exclusion protocol then checking robots.txt file is a better method for detection also if all robot's updated IP and user agents are available then techniques *IP and User agent mapping* are also better for detection.

Since every technique has an exception hence there is need to apply some other heuristic technique to detect robots. For Banaras Hindu Web site server log, combined technique is suitable to detect robots

Data sets	Duration	Total no. of requests
Dateset1	24/03/2014:00:00:0 to 24/03/2014:23:59:59	1266390
Dateset2	24/03/2014:00:00:0 to 26/03/2014:23:59:59	3248595
Dateset3	24/03/2014:00:00:0 to 30/03/2014:23:59:59	6005814

Table 2: Description of datasets extracted by Data Extraction algorithm

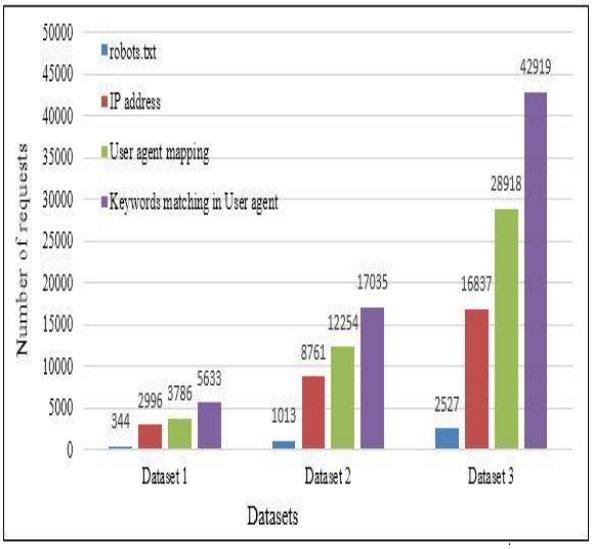

Figure 2: Graph of Robot detection techniques for different datasets

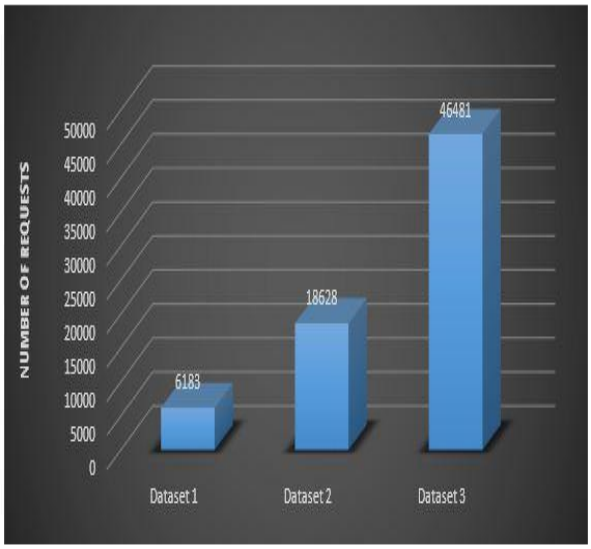

Figure 3: Number of Requests of robots by using combined technique

V. CONCLUSION

Web server log is a rich source of information which is used to predict user's navigation behavior. Due to exponential growth of information on Web, larger part of this log is filled by robot's requests. Sometimes it is necessary to detect robot's request for business organizations, Web usage analyst and web site administrator to protect their privacy, to distinguish robot from human user, to improve performance of server respectively. There are several techniques to identify robots in server log are robots.txt check, using IP address, User agent mapping, keywords matching in User agent etc. In this paper we have implemented these four techniques and combination of these techniques on Banaras Hindu University (BHU) web server log. After analyzing the results we conclude that combination of above four techniques works better for server log data in case of BHU Web server log.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

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] 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.
[3] Lu, Wei-Zhou, and Shun-zheng Yu. "Web robot detection based on hidden Markov model." 2006 International Conference on Communications, Circuits and Systems. 2006.
[4] Doran, Derek, and Swapna S. Gokhale. "Web robot detection techniques: overview and limitations." Data Mining and Knowledge Discovery 22.1-2 (2011): 183-210.
[5] Sardar, Tanvir Habib, and Zohreh Ansari. "Detection and confirmation of web robot requests for cleaning the voluminous web log data." IMpact of E-Technology on US (IMPETUS), 2014 International Conference on the. IEEE, 2014.
[6] Srivastava, Mitali, Rakhi Garg, and P. K. Mishra. "Preprocessing techniques in web usage mining: A survey." International Journal of Computer Applications 97.18 (2014).
[7] 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.
[8] Koster, M. 1994a. Guidelines for robot writers. http://info.webcrawler.com/mak/projects/robots/guidelines.html
[9] http://www.iplists.com/