International Journal of Research In Science & Engineering e-ISSN: 2394-8299
Volume: 3 Issue: 5 Sep.-Oct. 2017 p-ISSN: 2394-8280

"COMPARISON BETWEEN WEB ROBOT REQUEST DETECTION TECHNIQUES ON WEB SERVER LOG IN DATA MINING"

Nitika Kadam

Asst. Prof., Malwa Institute of Science & Technology, Indore
Kadamnitika01@gmail.com

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

Data mining is the computational process of discovering patterns in large amount data sets involving methods at the intersection of artificial intelligence, machine learning of Data System. The World Wide Web is now a huge database with this growth there arises a need for analyzing the data. The process of discovery and analysis of Web is called Web mining. Web mining is the application of data mining techniques to discover patterns from the Web.

Web mining can be divided into three different types

- Web Structure Mining:- Web structure mining is the process of discovering the connection between web pages.
- Web Content Mining:- Web content mining includes mining, extraction and integration of useful data and knowledge of Web page content.
- Web Usage Mining:- Web Usage Mining is a technique of extracting useful information from the Web Log, e.g. the pattern in which a user goes through different Web pages.

Mining enterprise proxy log plays an important role for enterprise manager and employer which makes it difficult to find the "right" or "interesting" information [1]. Web Log are generally noisy and ambiguous. Web applications are increasing at an enormous speed and its users, are increasing at exponential speed.

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 site administrator to solve maintenance issues like checking the broken hyperlinks and mirroring can also use web robots. 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 want 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 behavior.
- 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].

II. LITERATURE SURVEY

The following section discusses the various works of several authors.

Zheng L.et al [4] has proposed Optimized User Identification, Optimized Session Identification. The optimized data preprocessing technology is used to improvement of the technology betters the quality of data preprocessing results. The strategy based on the referred web page is adopted at the stage of user identification. Experiments have proved that advanced data preprocessing technology can enhance the quality of data preprocessing results.

Munk M.et al [5] has tried to assess the impact of reconstruction of the activities of a web visitor on the quantity and quality of the extracted rules which represent the web user behavior patterns. Experiment, find out to which criteria are necessary to realize this time-consuming data preparation and specifying the inevitable steps that are required for obtaining valid data from the log file.

Tyagi N.et al [3] provides an algorithmic approach to data preprocessing in web usage mining. They take requests for graphical page content, or the other file which can be induced into a web page, or navigation sessions performed by robots and web spiders into consideration.

III. WEB USAGE MINING

Web Usage Mining could be a technique of extracting useful information from the web log, e.g. the pattern in which a user goes through different Web Pages.

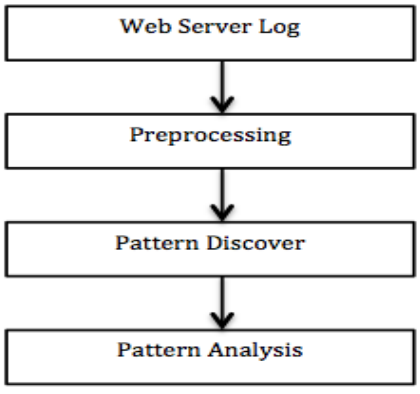

Figure 1. Process of Web Usage Mining

IV PROPOSED WORK

In this work we compare methods, which identify web robot request from web server, log file. Methods are

- User-Agent Check
- IP Address
- Head Count
- Robot.txt Access
- Hybrid Method

User Agent Base Method: Instance, the user agent field of Web robots should contain the name of the robot, unlike the user agent field of Web browsers, which often contains the name Mozilla as shown in table III. In this method User-agents field of the log file is verified. If user agent field value matches then it is a web robot.

IP Address Check: Another way to detect robots is by matching the IP address of a client against those of known robots. There are many Web sites that provide a list of IP addresses for known Web robots. Many web sites available, which provides up-to-date IP addresses of the web robot clients. As the same IP address could be used by Web users for surfing the Web and by robots to automatically download some files from a Web site, some other method should also be used along with this method to get confirmed about the robot request. Selecting only intersection of resulting IP addresses can do this.

Count of HEAD requests: The guidelines for Web robot designers also suggest that ethical robots should use the HEAD request method, whenever possible. The request method (e.g. GET, HEAD and POST) of an HTTP request message decides what type of job the Web-server should execute on the resource requested by the Web client.

Robots.txt Accesses: It is a file kept in the top-level directory of a web server. When a robot searches for the "/robots.txt" file for URL, it removes the path section from the URL and puts "/robots.txt" in its place. For example, for “http://www.anysite.com/searchany/index.html, it will remove the ""searchany/index.html ", and replace it with "/robots.txt", and will end up with http://www.anysite.com/robots.txt.

Hybrid Method: We start with the log file as the source of our experiment. Next we have gone through pre-processing. In this stage, Requests containing image access requests are deleted and query strings are eliminated. Image access requests of web robots are very rare. Query strings are eliminated to shorten the job of searching the voluminous log file entries. After this, using the pre-processed log file each of our four methods are executed separately named M1, M2, M3, M4 in the figure respectively for robots.txt check, user-agent check, IP address check, Count of HEAD requests and HTTP requests with unassigned referrers. Output request set obtained by robots.txt checking are confirmed as web robot requests.

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.

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] N. Tyagi, A. Solanki, S. Tyagi, "An Algorithmic Approach to Data Preprocessing in Web Usage Mining", International Journal of Information Technology and Knowledge Management 2 (2) (2010) 279–283.

[3] Ling Zheng Hui Gui. Feng Li, "Optimized Data Preprocessing Technology for Web Log Mining",International Conference On Computer Design And Applications (ICCDA 2010).

[4] Michal Munk, Jozef Kapustaa, Peter Šveca*, "Data Preprocessing Evaluation for Web Log Mining:Reconstruction of Activities of a Web Visitor" International Conference on Computational Science, ICCS 2011 P5ocedia Computer Science 1 (2012).

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] Koster, M. 1994a. Guidelines for robot writers. Htp://info.webcrawler.com/mak/projects/robots/guidelines.html