Identifying Bot Traffic & Analyzing Frequently Accessed Web Pages Using Web Log Files

Shouaib Mujawar[1],Anugrah Singh[2],Daksha Vighne[3]

Prof. Pratima Patil

B.E., Dept. of Computer Engineering, Trinity Academy of Engineering, Pune, Maharashtra,India

Abstract--*Web mining could be a term accustomed discover and extract helpful information from the world Wide net. It involves the automated discovery of patterns from one or additional net servers. An entire preprocessing technique is being planned to preprocess the web log file for extraction of user accessibility patterns .Data cleanup algorithmic rule removes the unrelated records from journal and filtering algorithmic rule discards the uninterested attributes from log file. It's the method to extract the user sessions from the given log files. Initially, every user is known consistent with his/her IP address per the log file and corresponding user sessions ar extracted. Different kinds of logs i.e. server-side logs and client-side logs ar normally used for web usage and usefulness analysis. Recent business reports assure the increase of web robots that comprise over half the overall net traffic. They not solely threaten the safety, privacy and potency of the online however they additionally distort analytics and metrics, ruining the truthfulness of the data being promoted. Hackers also use bots to threaten website admins by engaging thousands of bots on their websiten which casuses Denial of Service attacks. Usability is outlined because the satisfaction, potency and effectiveness with that specific users will complete specific tasks in a very explicit setting. This method includes three stages, specifically knowledge cleanup, User identification, Session identification. We will implementing these 3 phases relying upon the frequency of users visiting every page and then identify bots using a bot identification algorithm.. FP growth algorithm will be used to discover frequent patterns in the web logs such as frequently accessed webpages on a particular website.*

KeyWords--Web log mining, User identification, Session identification, Bot Detection, Frequently Accessed Pages.

I. INTRODUCTION

The web as a largest information construct has had abundant progress since its advent. Because the web has become customary these days, highlighted content generated by users, ability and usefulness. An online server could permit users to act and collaborate with one another in a very social dialogue as creators of user-generated context in a very virtual community. So, World Wide web becomes a standard and user friendly for transferring info. Therefore, individuals are more interested in analyzing log files which may supply additional helpful insight into web usage. Data processing is the extraction of data from the huge quantity of knowledge sets, to seek out relationships and patterns in data that are not antecedently been discovered to help the users. Web mining is one of the techniques of data mining to extract helpful information to supported users' desires. After web mining, web usage mining is one among the applications of data mining technology to extract information from weblogs to investigate the user access to websites. Web mining is the use of data mining techniques to discover and extract information from internet documents and services. There are three general categories of data which can

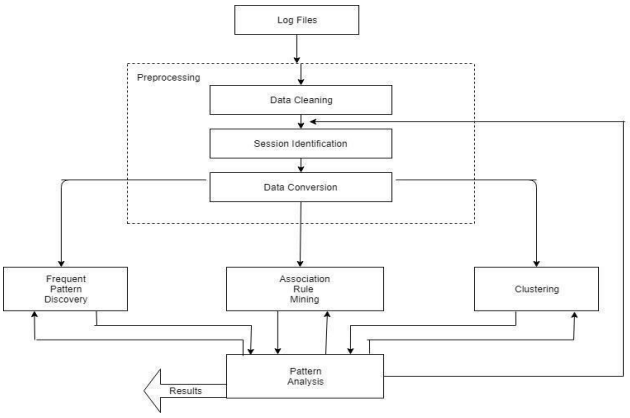
be discovered by web mining. Web activity: From server logs and application activity. Net graph: from links between pages, individuals and alternative knowledge. Net content: For the information found an online pages and inside documents. Web log files are the files that contain complete info regarding the users browse activities on the online server . These Log files ar created by each user click to the corresponding net servers. These log files are in text format and therefore the size varies from 1KB to one hundred MB. There ar 3 kinds of log files that are as follows: Web Server Logs, Proxy Server Logs, Browser Logs.

The advances in technology created potential to do activities from home like on-line transactions with banks, searching from on-line sites and on-line video games. This type of advance aims to comfort the web users. However, these technological enhancements ar encountered with a true threat appeared from criminals within the cyber world, UN agency exploit vulnerabilities for his or her malicious intents. Malwares and Malicious programs ar a sort of application that's designed and deployed for the aim of serving the trespasser in his/her malicious objectives. Worms, Trojans, viruses, backdoors, key loggers, botnets and ransom wares ar samples of malwares. Attackers use botnets to infect machines then manage them. Such approach for dominant associate degree infected machine is achieved by a communication theme named Command & management (C&C) channels. Internet robots, additionally called net crawlers, are pc programs that request resources from net servers across the web without human intervention. Constant growth of internet technologies and social media generate a large quantity of valuable info which can be accessed by net crawlers and rising advanced robots representing web of things devices, like good watches, cars and digital assistants. As of 2016, internet traffic originated by web bots constitutes over 51.8% of the overall net traffic. Malicious bots threaten the safety, privacy and performance of an online application. Non malicious bots ar concerned in analytics skewing, moving the dependableness of metrics and, by extension, the choice creating method. A recent business report points out that giant websites with distinctive content, like blogs, on-line newspapers and digital libraries of educational publications, ar the foremost engaging to bots.

II. LITERATURE SURVEY

In 2014,Grazyna Suchacka proposed a research paper “Analysis of Aggregated Bot and Human Traffic on E-Commerce Site” .In this paper a significant volume of Web traffic nowadays can be related to robots. Although some of them, e.g., search engine crawlers, perform essential tasks on a website, others may be malicious and should be banned. This paper investigates the share of bot-generated traffic on an e-commerce site and studies differences in bots' and humans' session-based traffic by analysing data recorded in Web server log _les. Both kinds of sessions reveal different characteristics, including the session duration, the number of pages visited in session, the number of requests, the volume of data transferred, the mean time per page and the number of images per page.

In 2016,Neha Singla 1 & Er. Navroz Kahlon2 proposed a research paper _Extracting Knowledge from Web Server Log Files and Future Prediction of Web Pages Using Kernel Based Fuzzy Clustering Method. In this paper useful information is summarized by analysing data from different interpretations and this process is known as Data mining.

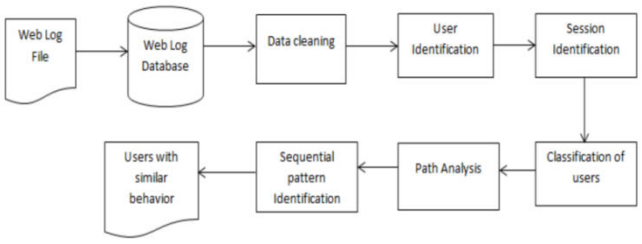


This can be used to cuts costs, increase revenue or both. There are several types of Data mining software or analytical tools for analysing data. Users use these tools in a different number of ways to analyse data.In this paper, a complete pre-

processing technique is being proposed to pre-process the web log file for extraction of user accessibility patterns. Data cleaning algorithm removes the unrelated records from web log and filtering algorithm discards the uninterested attributes from log file. From the desired attributes different patterns according to client users or accessed pages have been prepared. In the end future access has been predicted for a particular webpage by a particular IP address. This approach can be used for those webpages where there are similar access patterns.

In 2017,Grigorios Tsoumakas and Georgios Papadopoulos proposed a research paper "WebRobot Detection in Academic Publishing" .Recent industry reports assure the increase of web robots which composes more than half of the total web traffic. Web robots, also known as web crawlers, are computer programs that request resources from web servers across the Internet without human intervention .Malicious bots threaten the security, privacy and performance of a web application. Non-malicious bots are involved in analytics skewing, reacting the reliability of metrics and, by extension, the decision making process. In this paper, an approach on detecting web robots in academic publishing websites has been explained. Different supervised learning algorithms with a variety of characteristics deriving from both the log les of the server and the content served by the website have been used. The approach relies on the assumption that human users will be interested in certain domains or articles, while web robots crawl a web library incongruously. Experiments with features adopted in previous studies with the addition of novel semantic characteristics which derive are performing a semantic analysis using the Latent Dirichlet Allocation (LDA) algorithm.

In 2019,Sushmeendra N Rao, Rakesh B, Pallavi N Hegde, Anusha R kotur proposed a research paper _Predicting user behaviour through Sessions using the Web log mining for an E commerce application." In this paper users with similar behaviour are analysed. Weblog mining is the method to extract the user sessions from the given log files. Each user is identified according to his/her IP address specified in the log file and corresponding user sessions are extracted. Server-side logs and client-side logs are commonly used for web usage and usability analysis. Server-side logs are generated automatically by every user when users click to the corresponding web servers. Weblog mining process includes three process, namely data pre-processing, pattern analysis and pattern discovery. Data pre-processing includes 3 stages, namely Data cleaning, User identification and Session identification. In this paper, we are implementing these three processes and finding the users with same behaviour.



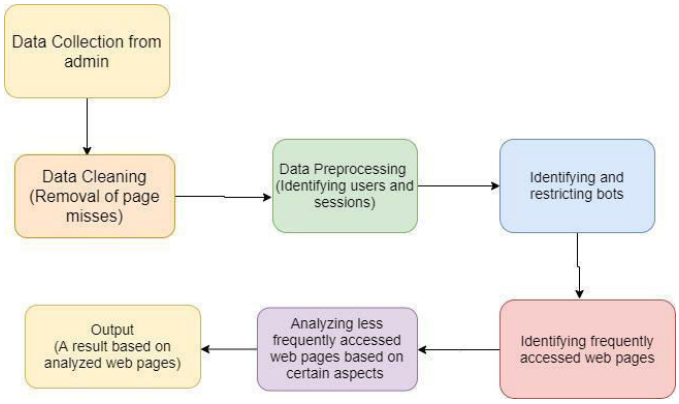
III. SYSTEM FLOW

1. Collect data from web server log files.
2. Data cleaning-Remove entries from web logs which are page misses.
3. Data Preprocessing- Identify sessions and users form the dataset.
4. Identify Bots from dataset using data mining techniques.

IV. SYSTEM ARCHITECTURE

We have proposed a system which will first detect the bots, remove them and then analyse frequently accessed web pages by the user using web log files .Firstly the data will be collected from the admin and then data cleaning will be done. Data cleaning is the process of detecting and correcting(removal) corrupt records from the data set and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Further data pre-processing steps will be implemented and then the system will identify the bots and remove them. the system will the apply the algorithms and analyse the web pages visited by the users. From this the less frequently accessed web

pages and frequently accessed web pages will be sorted and the output will be given respectively.



V. CONCLUSION

About 50% of the internets traffic comprises of bots. Bots can cause various attacks on your website such as Denial of Service and content scraping. Identifying bots is a must in order to protect your website as well as better for analysis of real users that visit the website. We
will be identifying bots by observing a certain type of bot behaviour. Frequently accessed webpages from a website are analysed using mining algorithms and these pages are analysed using certain page aspects such as page response time and quality of images and videos
present on the webpage.

VI. REFERENCES

[1] Neelima, G., & Rodda, S. (2016). Predicting user behavior through sessions using the web log mining. 2016 International Conference on Advances in Human Machine Interaction (HMI).

[2] Ruili Geng, and Jeff Tian "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage" IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 45, NO. 1, FEBRUARY 2015.

[3] G. Neelima and Sireesha Rodda, "An Overview on Web Usage Mining", Springer International Publishing Switzerland December 2015.

[4] S. Jagan, and S.P. Rajagopalan, "A survey on web personalization of web usage mining_, IRJETInternational Research Journal of Engineering and Technology, 2015.

[5] Phyu Thwe, "Using Markov Model and Popularity and Similarity-based Page Rank Algorithm for Web Page Access Prediction",ICAET'2014.

[6] Dilpreet Kaur1, A.P. Sukhpreet Kaur,"User Future Request Prediction Using KFCM in Web Usage Mining", IJARCCE, 2013.

[7] Mayank Kalbhor , Kunl Jain, "Fuzzy Based Hybrid Approach for User Request Prediction Using Markov Model",IEEE,2015.

[8] Gan TeckWei, Shirly Kho,Wahidah Husain, Zuraidah Zainol "A Study of Customer Behaviour ThroughWebMining" Volume 2, Issue 1 availableat www.scitecresearch.com/journals/index.php/2015.

[9] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu, "Web-Page Recommendation Based on Web Usage and DomainKnowledge", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 10, OCTOBER2014.

[10] Srivastva, Jaideep et al. (2000) _Discovery and Applications of Usage Patterns from Web Data_ ACM SIGKDD, Volume 1, Issue 2

[11] A. Ladekar, P. Pawar, D. Raikar and J. Chaudhari, "Web Log Based Analysis of User's Browsing Behavior", IJCSIT - International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015.

[12] A. Deepa, and P. Raajan, "An efficient preprocessing methodology of log _le for Web usage mining", NCRIIAMI - National Conference on Research Issues in Image Analysis and Mining Intelligence, 2015.

[13] N. Anand, "E_ective prediction of kid's behavior based on internet use", International Journal of Information and Computation Technology, 2014.

[14] A.P. Singh, R. C. Jain, "A Survey on Different Phases of Web Usage Mining for Anomaly User Behaviour Investigation", IJETTCS - International Journal of Emerging Trends & Technology in Computer Science, Vol 3, 2014.

[15] Meera Narvekar,Shaikh Sakina Banu, "Predicting User's Web Navigation Behaviour Using Hybrid Approach", ELSEVIER Science Direct,2015.

[16] Mamoun Awad and Issa Khalil ,"Prediction of User's web-browsing behavior Application of Markov Models" 2015, IEEE.

[17] Anshul Bhargav, Munish Bhargav , "Pattern Discovery and Users Classification through Web Usage Mining" 2014 IEEE.

[18] Priyanka S. Panchal,Prof. Urmi D. Agravat,"Hybrid Technique for User's Web Page Access Prediction based on Markov Model",4th ICCCNT 2013, IEEE.

[19] Virendra R. Rathod and Govind V Patel, "Prediction of user behaviour using web log mining in web usage mining", International journal of computer application vol. 139- No. 8, April 2016.

[20] Wei, Shirly Kho, Wahidah Husain, Zurinahni Zainol , "A study of customer behaviour through web mining ", Journal of Information Sciences and Computing Technologies ISSN 2394-9066 , Volume 2, Issue 1 February, 2015.