iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6306

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Determination of User Navigational Patterns from Server Log Files using Hadoop Techniques

Ruchi Patil[1], Prof. Puja Trivedi[2]
[1]*Research Scholar,* [2]*Assistant Professor, Department of Computer Science & Engineering, CIIT, Indore, M.P, India,*

Abstract: ***Web Usage Mining (WUM) is the major application of data mining to the web data and analyzes the user's visiting activities and attains their interests by inspecting the web log files. The first and crucial step in WUM is preprocessing of log for cleaning data, after that it can be suitable for mining purposes. The cleaning includes removal of unnecessary data such as image files, unclear data and robots. The aim of this research is to identify Robot Requests by analyzing web server log file using different methods such as IP address check, User agent check, Head request check, URL weight calculation and URL similarity calculation. Web Robots are the software programs that retrieve Web resources by traversing the hyperlink structure of Web. These robots can be used for malicious intensions; therefore they must be detected and removed from the log files. Eradication of robots aids to reduce log data and discern genuine users and there access patterns.***
Keywords: Web server, Web usage mining, web robots, web logs, URL similarity

I. INTRODUCTION

Number of websites and users are increasing gradually with the use of gradually increased in internet users. With this noteworthy increase of existing data on the Internet and because of its fast and disordered growth, web searching has become a tricky procedure for the majority of the users as it makes users feel confused and at times lost in overloaded data that persevere to enlarge. E-business and web marketing are quickly developing and to predict the requirement of their customers is obvious particularly. As a result, guessing the users' interests for improving the usability of web has turns out to be very essential [1].
Web pages contain huge amount of information that may not be concerned to the user. The major source of information regarding the users visited links, browsing patterns, time spent on a particular page or link, are Web Log Data. This information can be used diverse applications like adaptive websites, modified services, customer summary, generate attractive web sites etc[1]. The first and crucial step in WUM is preprocessing of log for cleaning data, to make it suitable for mining purposes.

A. Motivation

There are full of dynamic content in modern websites in the form of in-the-moment news articles, opinions, and social information. Consequently, Web robots or crawlers, which are software agents that automatically submit HTTP requests for content from the Web without any human interference, have been steadily growing in elegance and in volume. An industry report suggest that 6/10 of all the web requests are originated by robots, and this proportion is slated to rise to ever higher levels[3]. These robots are also deployed for malicious purposes such as Distributed Denial of Service attack, sending spam mails, they are able to perform human user tasks such as registering user accounts, searching/submitting contents etc.[4]. Web usage mining is a skillful and efficient way of extracting fruitful information (patterns of user accesses to a website) from the web logs. These robots can be distinguished and identified by analyzing web logs. Eradication of robots aids to reduce log data and discern genuine users and there access patterns.

B. Web Usage Mining

Web Usage Mining (WUM) is the process of extracting user access or navigational patterns by applying data mining techniques to the Web log files [5]. It contains three main steps data preprocessing, pattern discovery and pattern analysis. Figure 1, shows the process of WUM.

1) *Data Preprocessing:* This is most important phase in Web usage mining process. The data from the log file cannot be used directly for mining process [6]. Therefore the content of the log file must be cleaned in the preprocessing step. The different tasks of preprocessing are
2) *Data Cleaning:* The first step in data preprocessing is to clean the raw web data. During this step the available data are examined and irrelevant or redundant items such as records with filenames extension of GIF, JPEG,CSS. and noisy data are removed from dataset [7].
3) *User Identification:* For the success of a personalized website, the identification of individual users who access a website is one of the most important issues [7].

4) *Session Identification:* User sessions identification is very important as sessions encodes navigational behavior of users which is used for pattern discovery. This is done by using the time stamp details of the web pages. The total time used by each user of each webpage. Session is the time duration spent in the web page [6].
5) *Pattern Discovery:* Users are classified on the basis of their navigational behavior to discover knowledge or patterns [7].

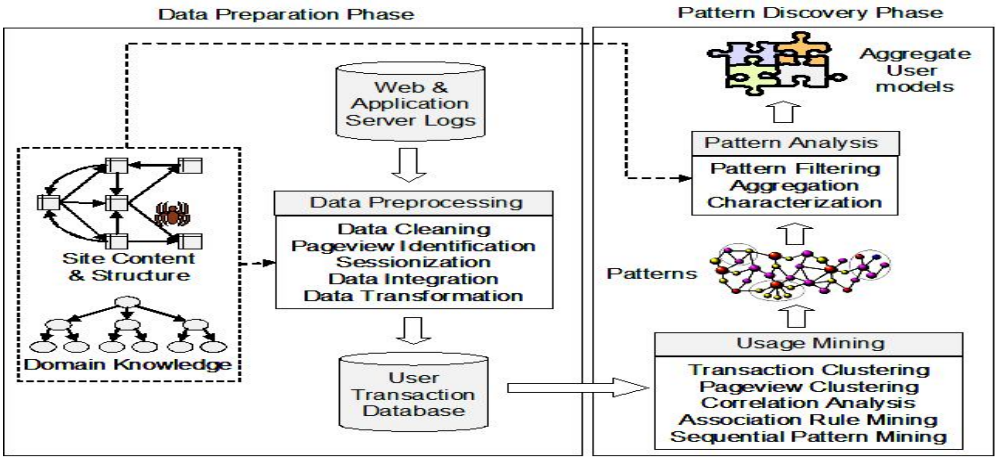


Fig 1. Process of Web Usage Mining

6) *Pattern Analysis:* WUM techniques were applied on the dataset and only the interested data is extracted from the discovered patterns and irrelevant data is removed [7].

C. Web Robots

Web Robots are the software programs that retrieve Web resources by traversing the hyperlink structure of Web [4]. At times Web robots are called as spiders, bots, crawlers or Web wanders. There are many purposes where a robot can be used such as [2]-

1) For resource discovery, crawling and indexing for search engines like Google, Yahoo etc.
2) As an offline browser which downloads some set of resources for browsing.
3) As a line checkers to check hyperlink validity.
4) As a shopping comparison robots to supervise, compare particular product prices on other e-commercial Web site.
5) As an email collector to collect record of emails provided on web page.
6) These can also be used by Web site administrator to solve maintenance issues like checking the broken hyperlinks and mirroring. Excluding these, there are some Web robots that are designed and employed for some malicious purposes like Distributed Denial of Service attack, sending spam mails, they are able to perform human user tasks such as registering user accounts, searching/submitting contents etc[4].

Need to Identify Web robots [8]

a) Business organizations on Web wants to collect business intelligence information by disabling unauthorized access of robots.
b) Web usage analysts/Researchers are keen to discern human user and robot to identify correct user's navigation behaviour. Eradication of robots aids to reduce log data and discern genuine users and there access patterns.
c) Sometimes larger part of network bandwidth is consumed by robots that slows down the speed of server response.
d) For privacy and security issues.

II. LITERATURE REVIEW

A wide variety of existing mechanism, algorithms and architectures is studied for identifying the issues removed and remains in Web Usage Mining. Later on, this will give a brief categorization of various approaches, which has been suggested over the last few years on detection and removal of web robot request.

Mitali Srivastava et al. [2] had studied about the Web robots that are present in the World Wide Web. They are the software programs that retrieve Web resources by traversing thehyperlink structure of Web. There are many ways where a robot can be used such as crawling and indexing information for search engines, offline browsing, shopping comparison and email collectors. Excluding these, Web robots can be used for some malicious purposes like DDoS, sending spam mails etc. Therefore it is necessary to detect them for privacy and security issues.

They had implemented various robot identification techniques such as robot.txt check, known robot's IP address check, User agent field mapping and key word matching in User agent
field. They had applied these techniques separately to the dataset and later they presented the combination of these four techniques.
Wang Dong et al. [11] had studied about various Web robot detection methods. The automated information gathering program (Web robots) that has brought many problems such as information leakage, resource occupation and network security threaten. Therefore it is important to control and detect them. They had proposed a new detection method with semi-supervised support vector machine. They had used feature extraction method to extract essential features from the log file via extracting access sessions to classify between humans and robots, then calculated the weight of extracted features, so as access logs can be converted to feature vector sets. These feature vector sets are then put into support vector machine to classify and identify Web robots.
Neha Goel et al. [12] had worked on preprocessing phase of Web logs in Web usage mining. Web usage mining is the process to find user access patterns from Web logs. Web logs contains highly unstructured data which is not suitable for data mining directly therefore preprocessing is done for making data suitable for analysis and also reducing file size. They had proposed a proper and absolute tool for preprocessing which removes irrelevant and noisy data and transform the data suitable for analysis. The tool is said to be absolute as after preprocessing and cleaning, it shows summary statistics of the records at the end. The summary statistics shows the number of records of input fed, after preprocessing the elements obtained and the computation time. At the end it exports .log file which contain the preprocessed data. This file can be imported in any data mining utility.
Ying Han et al. [13] had focused on the preprocessing phase in Web log mining. In Web data mining, Web log mining is the most important method, in which primary work is of data preprocessing. They had presented a method for data preprocessing based on user characteristic of interests. Records are filtered as low value records(low interest) and high value records(high interest) based on user interest and a threshold value. They had also compared users, based on their similarity. This records can be used to filter noise data, reduce the size of server data and based on similarity of user, this data can provide a model for recommendation system.
Mirghani. A. Eltahir et al. [7] had worked on Web usage mining technique to extract knowledge from unstructured data. In order to improve the performance of a website, it is important to understand users and learn about them from registered usage information in log files. They had proposed Web usage mining techniques to procure the knowledge from Web server log files. Web usage mining contains three steps for discovering knowledge that are preprocessing, pattern discovery and pattern analysis. In preprocessing phase, unwanted data is removed by applying several preprocessing techniques. In the next phase that is pattern discovery, users are classified on the basis of their navigational behavior to discover knowledge or patterns. In the last phase, pattern analysis, only the interested data is extracted from the discovered patterns and irrelevant data is removed. The WUM techniques were applied on the dataset and knowledge extracted on the factors such as top visited sites, popular paths through site, visitors stay length, visitors by hour of day, top search engines, popular path through sites and server logs.
Shinil Kwon et al.[10] had studied that, with the growth of Web based economy, it has become very important to detect Web robots. In order to differentiate between humans and Web robots, it is important to find features that are common characteristics of diverse robots. They had proposed a fresh approach that expresses the behavior of interactive users and various Web robots in terms of a sequence of request types, called request patterns. They had targeted to the detection Web robots such as text crawlers, image crawlers, email collectors and link checkers. An evaluation is done on more than 1 billion requests collected at www.microsoft.com and 94% accuracy achieved in detecting Web robots, estimated by F-measure. They had compared there approach with the decision tree algorithm and found there approach is more accurate.

III. PROBLEM IDENTIFICATION

Wide variety of existing mechanism, algorithms and architectures is studied for identifying the issues removed and remains in web usage mining. Later on, this gives a brief categorization of various approaches, which has been suggested over the last few years on detection and removal of web robot request. Some problems that need to be resolved for web server log analysis is listed below:

A. Infected system spreading malware
B. Compromised system
C. Successful attack:
D. Insider abuse and intellectual property theft:
E. Covert channel/hidden backdoor communication
F. Increase in probing
G. System crash

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

IV. PROPOSED METHODOLOGY

The architecture of proposed method consists of following steps:

1) *Preprocessing:* In this, data cleaning is performed where the data which are not useful are removed from the log file such as records with image file extensions(.jpeg, .gif, .png etc.) and noisy data.
2) *Tree Generation and Weight Calculation***:** In this method, tree is generated and on the basis of generated tree, weight of each URL is calculated on the basis of the fields such as Client IP address, Client URI, Client Method, User Agent, Referrer
3) *Weight Comparison:* A weight threshold is set to compare the URL weight obtained. Therefore the URL weights greater than the weight threshold is identified as robot request.
4) *URL Similarity Calculation:* In this method, URL similarity is calculated and similarity score is found by using a method known as Levenshtein Distance. Levenshtein Distance is a measure to calculate the similarity between two strings.
5) *Similarity Score Comparison:* A similarity threshold is set to compare the similarity score of URL. Therefore, the similarity score less than similarity threshold is passed to next step for identification.
6) *Matching URL with robot.txt file:* In this, the resulted URL from the similarity score comparison is matched with a file known as robot.txt. If URL matches with the URL in robot.txt file then the URL is identified as robot request.
7) *Robot.txt file:* It is the file that contains the URL of known robots.
8) *Merging Results of both methods:* In this method, the URL identified as robot request from both the methods that are Tree Generation and Weight calculation and URL similarity calculation are merged and confirmed as robot request.

The flow chart of proposed method is shown in figure 2.

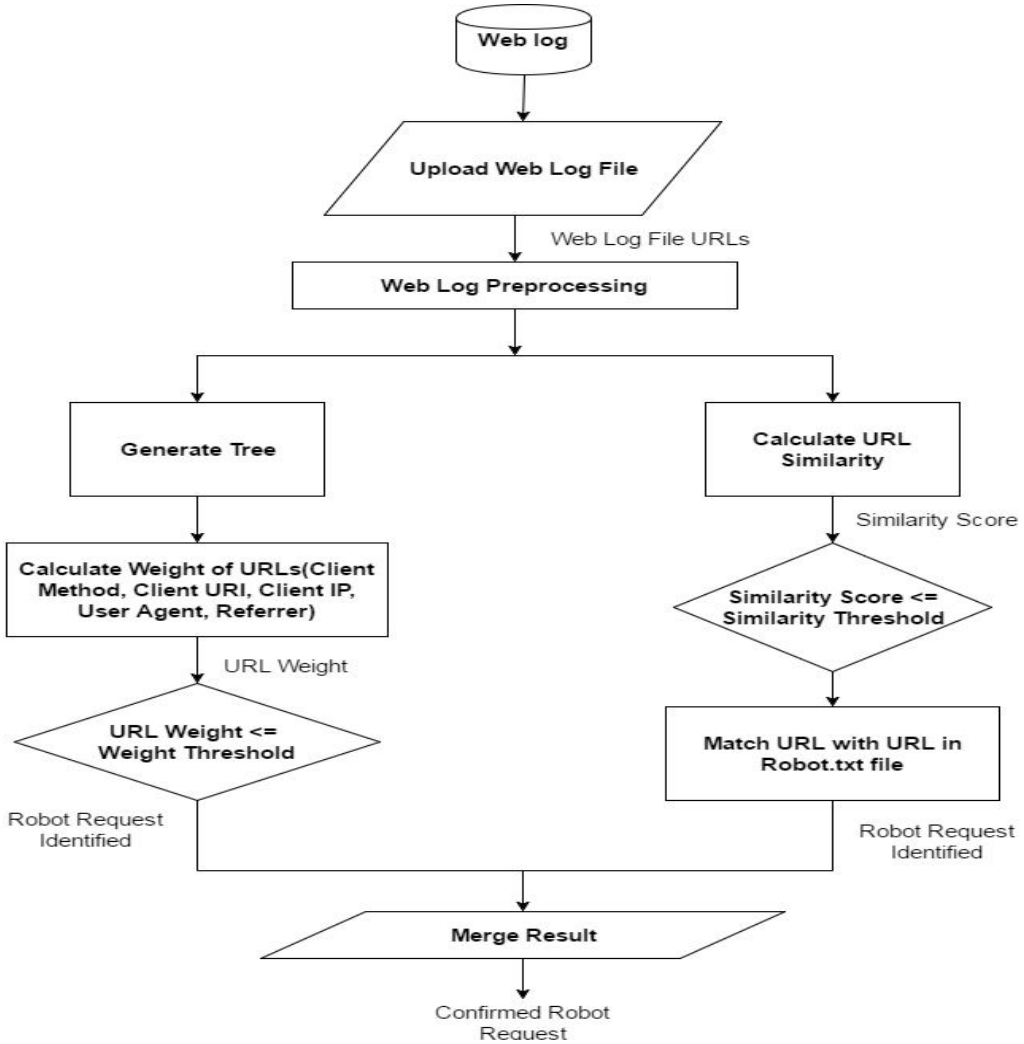


Fig. 2 Flow chart of proposed method

V. RESULT ANALYSIS

Proposed system implemented with JDK1.8, Apache Tomcat Application Server, JFree Chart Library as the software configurations. The following evaluation parameters are used

A. Computation Time
B. Number of Robot Request using
1) User Agent
2) IP Address
3) Head Request

C. Number of Robot Request using
1) Tree Generation and Weight Calculation

The proposed approach was tested on four data sets, Table 1 shows the comparison of total no of Robot request Identified in existing system and proposed system applied on four datasets.

TABLE I
EXPERIMENT PERFORMED ON THE DATASETS FOR NO OF ROBOT REQUEST

Number of entries in log files	Number of robot requests	
	Existing System	Proposed System
150	67	84
220	78	137
500	199	213
1000	354	402

Figure 3 shows the graph for comparison of Total no of Robot request Identified in existing system and proposed system applied on four datasets.

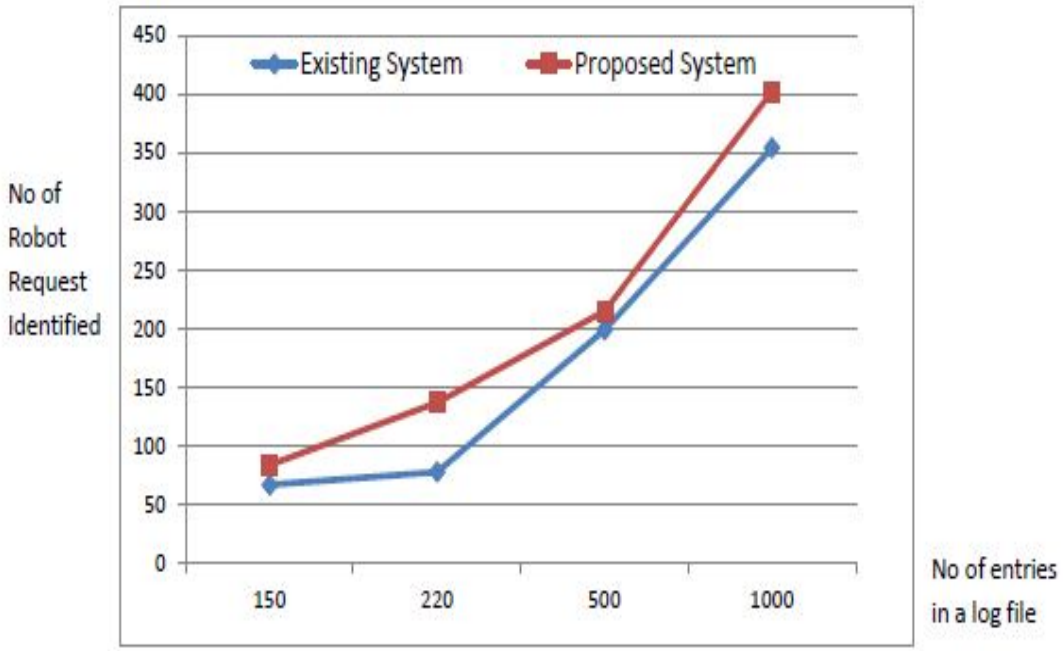


Fig 3 Result set comparison of Dataset (No of robot request)

Table 2 shows the comparison of computation time found in existing system and proposed system applied on four datasets.

TABLE 2
EXPERIMENT PERFORMED ON THE DATASETS FOR COMPUTATION TIME

Total Number of entries in log files	Computation Time (in ms)	
	Existing System	Proposed System
150	3.94	3.0
220	5.2	4.0
500	7.4	7.0
1000	12.67	10.4

Figure 4 shows the graph for comparison of computation time found in existing system and proposed system applied on four datasets.

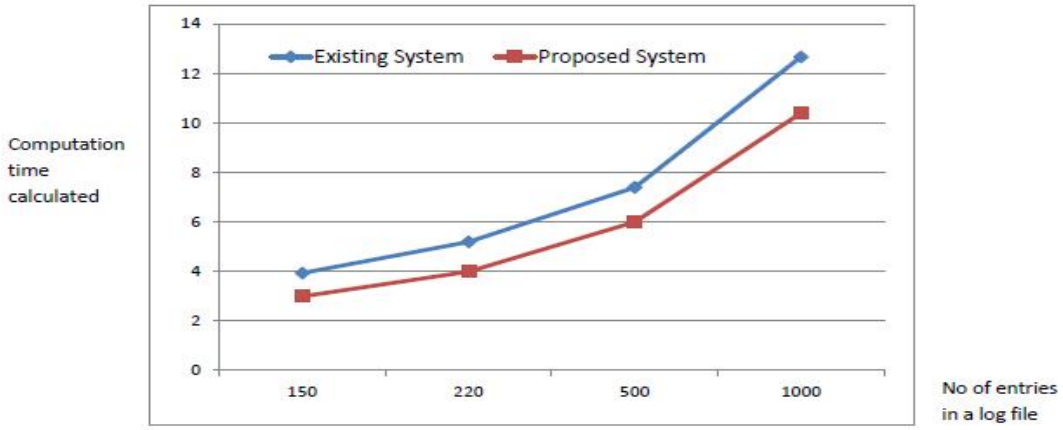


Fig 4 Result set comparison of Dataset (Computation Time)

Developing a solution is an approach proving mechanism but to prove its results is a complicated task because it measures each and every step of the solution and let it compare with the existing mechanisms. The results are going to be verified on the basis of the analysis. Only few results has been discussed here.

VI. CONCLUSION

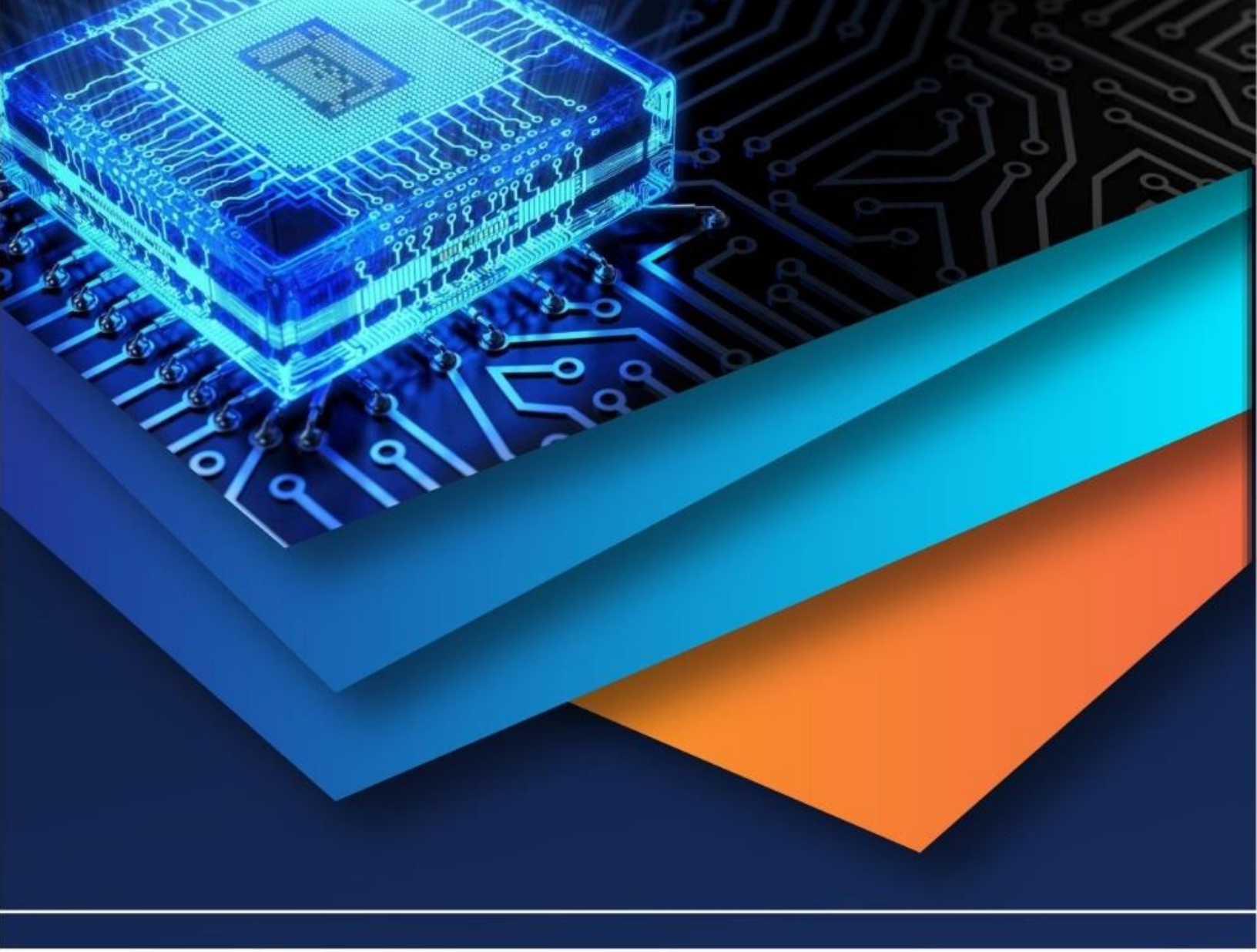
WUM technique is very good to extract knowledge from unstructured data. The web administrator or the web designer can use the obtained results of WUM to organize their website by determining system errors, user's preferences, technical information about users, and corrupted and broken links. The detail of methodology is presented that utilizes and integrates four different robot request detection methods and integrates them together to detect and finally verify requests as a confirmed robot request. The integration is achieved by utilizing Set Union and Intersection operations.

Two different methods are proposed to identify robot request: Tree Generation and Weight Calculation and URL Similarity to identify a robot request. The result set from both the methods are merged to confirm a robot request. The proposed approach is more reliable and takes less computation time.

REFERENCES

[1] P.Nithya, Dr.P.Sumathi, "Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots", 2012 National Conference on Computing and Communication Systems (NCCCS) , 978-1-4673-1953-9/12 ©2012 IEEE.
[2] Mitali Srivastava, Atul Kumar Srivastava, Rakhi Garg, P. K. Mishra, "Comparative Analysis of Robot Detection Techniques on Web Server Log", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2015.
[3] http://wiki.knoesis.org/index.php/Understandingand_Mitigating_the_Impact_of_Web_Robot_and_IoT_Traffic_on_Web_Systems.
[4] Tanvir Habib Sardar, Zahid Ansari, "Detection and Confirmation of Web Robot Requests for Cleaning the Voluminous Web Log Data", 2014 International Conference on the IMpact of ETechnology on US (IMPETUS), 978-9332-9026-40/14 ©2014 IEEE.

[5] Ms Shashi Sahu, Asst Prof. Leena Sahu, "A Survey on Frequent Web Page Mining with Improving Data Quality of Log Cleaner", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 3, March 2015.
[6] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai, "Analysis Of Web Logs And Web User In Web Mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.
[7] Mirghani. A. Eltahir, Anour F.A. Dafa-Alla, " Extracting Knowledge from Web Server Logs Using Web Usage Mining", 2013 International Conference On Computing, Electrical And Electronic Engineering (ICCEEE), 978-1-4673-6232-0/13 © 2013 IEEE.
[8] 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.
[9] Shinil Kwon, Myeongjin Oh, Dukyun Kim, Junsup Lee, Young-Gab Kim, Sungdeok Cha, "Web Robot Detection based on Monotonous Behavior", © Springer-Verlag Berlin Heidelberg 2012.
[10] Shinil Kwon, Young-Gab Kim, Sungdeok Cha, " Web robot detection based on pattern matching technique", Journal of Information Science, 38 (2) 2012, pp. 118–126 ©The Author(s), DOI: 10.1177/0165551511435969.
[11] Wang Dong, Xi Lei, Zhang Hui, Liu Hebing, Zhang Hao, Song Ting, " Web robot detection with semi-supervised learning method", 3rd International Conference on Material, Mechanical and Manufacturing Engineering (IC3ME 2015) © 2015. The authors - Published by AtlantisPress.
[12] Neha Goel, Dr. C.K.Jha, "Preprocessing Web Logs: A Critical Phase In Web Usage Mining", 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India, 978-1-4673-6911-4/15 ©2015 IEEE.
[13] Ying Han, Kejian Xia, "Data preprocessing method based on user characteristic of interests for Web log mining", 978-1-4799-6575-5/14 © 2014 IEEE, DOI 10.1109/IMCCC.2014.182.
[14] Theint Theint Aye, "Web Log Cleaning for Mining of Web Usage Patterns", 978-1-61284-840-2/11©2011 IEEE.
[15] http://www.robotstxt.org/robotstxt.html.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY