RESEARCHERID
THOMSON REUTERS

[Kiruthika, 6(5): May 2019] ISSN 2348 – 8034
DOI- 10.5281/zenodo.3174525 Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

WEB SERVER LOG ANALYSIS USING WEB USAGE MINING

V. Kiruthika*[1] & Dr. P. Pandi Selvi[2]

[*1]M.Sc Computer Science, Dr. Umayal Ramanathan College for Women, karaikudi

[2]Assistant Professor, Department of Computer Science, Dr.Umayal Ramanathan College for Women, Karaikudi

ABSTRACT

Log Files are the data files which include the log information like the name of the user, Date and access time, request type, transferred bytes from server and client, response status and accessed URL. These files are created and maintained by web servers as well as server tools like IIS, Apache and etc. It's necessary to know the interest of web users to offer effective services. The main theme of analyzing web usages with the help of log file helps to understand the users future needs and maintain the content of the website based on the previous data. This paper discuss about the log files and its management. The paper also proposes a graphical representation of web usage results.

I. INTRODUCTION

Web technology is working with the internal approach called client server technology. A Machine or computer which can deliver web pages is called web server. The web server is well structured and has an internal storage to hold web data. Simultaneously the server maintains background files to display the pages to the client computers. When the request comes from an IP address, the server receives the request details from the client and performs the task to send response. The server contains the images, audio files, video files and hypertexts internally and it combines all in a readable way and delivers as the response to the requested client. All the above media are converted into the hypertext and the client can get the entire media details as a response from server. The same procedure will be followed for every client based on the request. The log file is the data file which can hold all the transacted information from the client to server and server to client. The log file resides in the server machine.

The main objectives of the proposed work are highlighted below:

1. To propose a new way of web usage mining processes.
2. To provide faster performance. The Proposed approach is designed to load and process the web server log file in a fast and efficient manner.
3. To ensure trustworthy. The proposed research work exactly loads the content of log file and shows a preview to the user. The raw data is now converted into readable data.
4. To provide an enhanced unique framework to analyze the web server log files.

II. LITERATURE REVIEW

This section focuses on some of the related works that has already been done in this area:
K.Joshila Grace and V.Maheswari [6]., performed a brief investigation on weblogs. Their work provides an idea of creating an extended log file. In order to learn the user behavior, they used the technique of Weblog software tool on weblog datasets to extract the fruitful information by performing deep analysis. But, comparative study of weblog analysis may not done.

S.Padmaja, et al [8]., investigate the web server logs using log analyzer tool. They analyzed the log datasets by using Apache hadoop. Their approach will solve all issues caused by the existing systems. Based on their approach performance also increases. But their method can be used for business only.

Mayor Mahajan,et al [9]., investigated about Real-time weblog analysis. These are the automated and systematic process towards the analysis of emerging information and users' response. Their method is applicable for tables and reports only.

Kenneth J Klassen ,Waynesmith [14] briefly discussed about the weblog analysis. Their approach provides an idea of creating an extended log file and learning the user behavior. Their paper gave a detailed discussion about log files their format of creation, access procedures and their uses.

Lavanya ks, srinivas R[10] introduced web server logs using hadoop framework. Their system analyzed the log files and presented it to the user in a more understandable form such as piecharts, barcharts and so on. The fluctuations in their results cannot be shown as reports in online as per the changes of log.

III. PROPOSED METHOD

The proposed method analyses the web server logs with the help of Deep log Analyzer program. The analyzed results can be seen as, knowledge to answer these questions. How to extract knowledge from incomplete data structure? Is the log data that is gathered about the users is enough to understand them? What is the optimal structure and content of a web site in order to attract the maximum interest of visitors? What does the user want to do? What is the suitable method and technique of web mining to extract knowledge? Many different types of results occur according to WUM technique used. Visualization of WUM results should be expressed in high-level languages. The knowledge can be easily understandable and usable to humans. The obtained results of WUM can be used by web administrator or web designer to organize their website by determining system errors, user's preferences, technical information about users, and corrupted and broken links.

Web mining is the task and process of mining and extraction of data from the documents available in World Wide Web. Web mining is a part of Data mining. The web mining is done by the various techniques to extract the data from a large data set on the World Wide Web.

The proposed system is an online log analyzer which provides one time free registration. Any user can register with the web portal and do a remote login to access the services. The user can upload the log file into the web portal after the remote login. The website converts the raw data into readable dataset and performs a structured query mining techniques to filter relevant data. Once the data are filtered the website allow the user to analyze the data and generate dynamic charts.

IV. PARTS OF LOG FILE

Log file parts, as well as the contents of the log file, may differ and the contents are decided by the server. Basically the log file may contain the following information:
Name of the User: The name of the user is the client IP address and also the name of the user who had visited the web page. The IP address is retrieved from the internet service provider. The IP address is the important data for getting the client profile and access. This can help the server to identify the new visitor and the existing visitor.

Date and Time: The Date and Time data are important to know the access frequency. This field maintains the date of access and the time. Every access will be maintained as a new entry in the log file. So the date and time can offer the server to know the access frequency of the current online user.

Server IP: The server IP address is obtained and maintained in this area. A server may be spoofed and delivers the pages from various IP addresses. In this server IP field, it can maintain the IP address of the server which served the web page to the client.

Server Port: The server maintains the present port number of the server. The port number may be varying based on the availability of ports, when a client is requesting pages or server data.

Bytes Sent: The value of sent data is calculated in bytes format and the total sent bytes are maintained under this field. The sent bytes are generated for every access and for every log. **Bytes Received:** The sent byte represents the download bytes and the received bytes represent the uploaded bytes. This field can maintain the transferred data

from client to server. The total received bytes are calculated and maintained under this field. The sent bytes are generated for every access and for every log.

Bytes Received: The sent byte represents the downloaded bytes and the received bytes represent the uploaded bytes. This field can maintain the transferred data from client to server. The total received bytes are calculated and maintained under this title.

These are the basic and fundamental data present on a log file which is created and maintained by server machine. With the help of this log file the server or website administrators can get the access and request histories.

The proposed approach includes the following major processing units:

Data preprocessing

The log file data are called as raw data and the raw data can't be utilized for data mining. So it is necessary to convert the raw data into readable data. These conversion tasks are performed in the data preprocessing stage. The Datacleaning stage is performed at first to clean the unwanted data records. For example, The Server log file creates the bookmark of the current server software and the version of the software. Also the log file shows the field headings in the top row. The header rows and the row data are unnecessary. So, these kinds of data are eliminated during the data cleaning process.

Data detection

The data detection is the second task performed in the proposed method to discover the data in the server log files. The data are retrieved by using special mining based queries and the retrieved results are classified into categories. This stage can collect the required data like client IP, name of the client, sent and received bytes. The inbuilt code analyzes the data by the format of the data and the position of the data.

Data conversion

The data conversion is the final stage which helps to convert it into the readable format of data to perform mining methods. In the proposed work the data are converted and it will be maintained as a dataset with the help of SQL database.

Pattern discovery

The pattern discovery is the next stage and it will be performed after the successful completion of data preprocessing. By default the patterns are represented in a graph model. The patter discovery is done in this work by the association rules. The association rules are utilized to find the correlation of data as per the presence. The sequences of IP addresses are associated with each other and the IP address is filtered by eliminating the duplicates. Also the data rows are classified by the date and time. All the discovered data are grouped with the field headings.

Pattern analysis

The pattern analysis is the final stage. Analysis is the major task which can eliminate the irrelevant records from the required records. Pattern analysis is done in this work with the help of Structured Query Language (SQL).

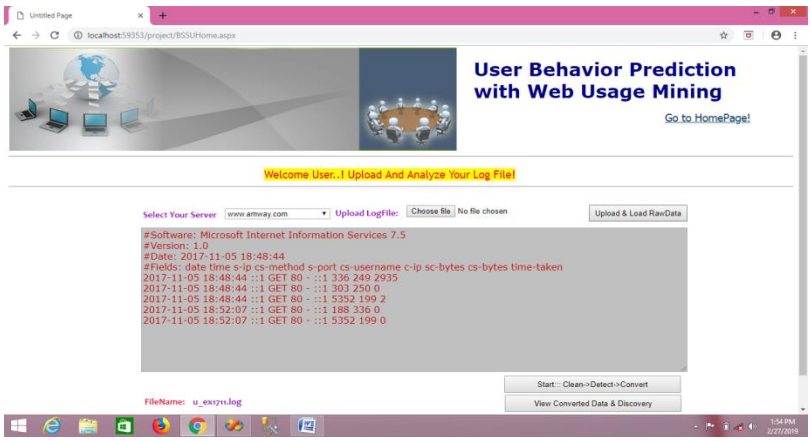
The server administrator can get the converted data from the proposed research web application. After the conversion process is completed the website redirects the analyzing page to the current online user. The website generates internal queries to retrieve the list of IP addresses which are accessed by the server. The internal code eliminates the duplicates and the unique data is shown to the online user. When the user select an IP address from the list, the web tool calculates the number of access counts, total sent bytes and total received bytes.

After the data are associated into a single dataset the web page generates dynamic chart reports to show the results to the online user as well as the server administrator. All these tasks and processes are performed internally by the proposed web application. The proposed research work does the above tasks by the association rules and Structured Query language.

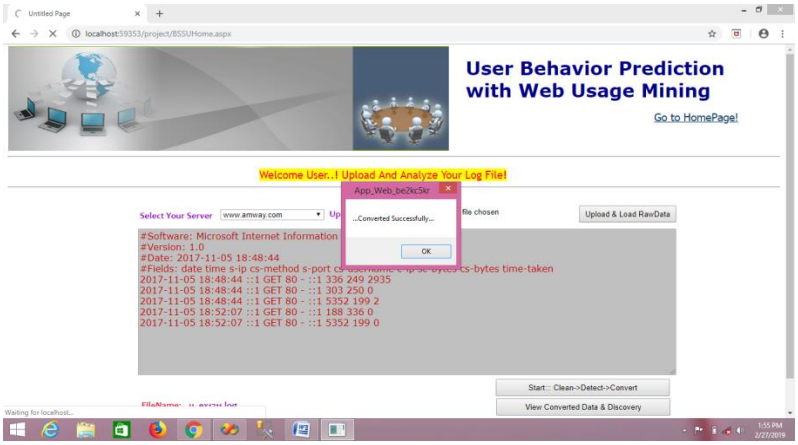
ISSN 2348 – 8034
Impact Factor- 5.070

V. EXPERIMENTAL RESULTS

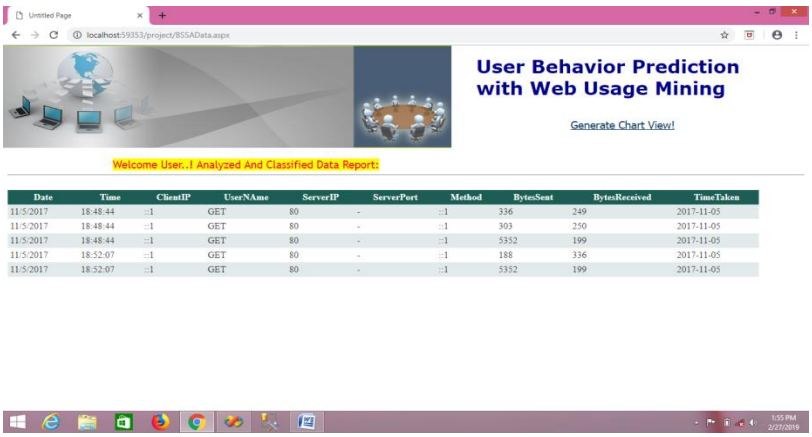
(i)Raw Data in a Log file:



(ii)Data clean, detection & conversion results:



(iii) Converted Data



VI. CONCLUSION

A web service can be more popular when it is offering solutions, based on the expectation of the user. But the identification and determination of user's expectation is more tedious and critical. But with the effective prediction system it will be easy and efficient. This research work proposes the method of easiest learning with the help of retrieved user's log file. Analyzing log file is the easiest way to identify and understand the user needs.

The approach proposes a minimized work of log analysis. The graphical representation of analyzed data can attract the administrators. The same method can be applied to any kind of web site such as Ecommerce, social networks and etc.

The internal codes and techniques are well formed to handle complex dataset and different fields. This research work offers a clear idea about log files and how to utilize the results for our needs.

REFERENCE

1. *S.Vijayalakshmi V.Mohan, S.Suresh Raja, (2009) "Mining Constraint-based Multidimensional Frequent Sequential Pattern in Web Logs," European Journal of Scientific Research., Vol.36.*
2. *D.Vasumathi, and A.Govardan,(June 2009) "BC-WASPT : Web Acess Sequential Pattern Tree Mining," IJCSNS International Journal of Computer Science and Network Security., Vol.9.*
3. *J. Dean, S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," In Proc. of the 6th Symposium on Operating*
4. *SystemsDesign and Implementation, San Francisco CA, Dec. 2004. journal of engineering science &research Technology, ResearcherID ,547, dec-2017*
5. *karuna p.joshi ,anupamjoshi,YELENA YESHA ,"on using a warehouse to analyze weblogs ",kluwer academic publishers,162,3003*
6. *L.KJoshila Grace ,v.Maheswari ,Dhinaharan ,nagamalai "Analysis of weblogs and web user in webmining",international journal of network security &its applications, 99, jan -2011.*
7. *Divyesh Bhior,Amey chavan ,Aditya patil, prof vaibhav pawar,"map reduced based log analysis and prediction using hadoop",international journal of scientific research engineering &technology ,volume 6,282 issue3, march-2017.*
8. *[8] S.padmaja,DR.Ananthi sheshasayee ,"web server logs to analyzing user behavior using log analyzer Tool",International journal of Advance research in science &engineering, 514, sep-2014.*
9. *Mayor mahajan,omkar Akolkar,Nikhilnagmule,sandeepRevanwar,"Realtime weblog analysis and hadoop for Data Analytics on large weblogs".International Research journal of engineering and technology",1827, march-2016.*
10. *Lavanya ks, srinivasa R, "customer behavior analysis of webserver logs using Hive in hadoop framework",International journal of Advanced networking &applications,409,2016.*
11. *DeeptiSahu,Shweta meena,"Detecting users behavior from web access logs with Automated Log Analyzer Tool",International journal of computer science and information Technologies ,5106,2014.*
12. *kavita Agrawal,Reader Hemant Makwana,"Data Analysis and reporting using Different Log Management Tools",International journal of computer science and mobile computing ,224,2015*
13. *Bernad J.jansen "Search log analysis:what tit is,what's been done how to do it",Library &information science ,ELSEVIERresearch 407,2006.*
14. *Kenneth J.Klassen,waynesmith,"weblog Analysis:A study of Instructor Evaluations Done online journal of information Technology Education 292,2004.*
15. *WEIXI,"Automatic log Analysis using Machine learning"Examensarbete 30hp,Uppsala universiTET,1,2013.*
16. *S.Alspaugh ,Beidichen,JessicaLin,Analyzing Log Analysis :An Emprical study of user Log mining "WW.SPLUNCK.COM 1,2014.*
17. *sak Taksa ,Amandaspink,Bernad J.Jansen,"Web log Analysis :Diversity of Research methodogies" copy@IGIGlobal distributing in print or electronic forms without written permission of IGSI Global is prohibited 504,2009.*

18. *Sonia Sharma,Munishwar Rai"Customer Behavior Analysis using Web Usage Mining",International Journal of scientific Research in computer science and Engineering,47,2017.*