IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

WEB SERVER LOG ANALYSIS

[1]A. Bamila Rachel, [2]S. Augusta, [3]A. Divyanandhi, [4]J. Fatima Stany, [5]M. Sowmiya

[1]*Faculty*, [2,3,4,5]*UG Scholar*

Computer Science and Engineering

Jayaraj Annapackiam CSI College of Engineering, Nazareth, India.

Abstract: An abstract for a web server log analysis would describe the major features and functionality of a web development logging solution. The goal of web server log analysis is to give developers an effective mechanism to log and track faults and other key events in their web applications. Log files have become a typical component of large applications and are required in operating systems. It may make it easier to identify unknown IP addresses that assault the system.

Index Terms **- Log files, web, server**

I. INTRODUCTION

Web-based apps have become commonplace in today's globally connected environment. The log files in our project include a large number of log entries. Entries such as IP addresses, OS system type, origin of IP address, and threat associated with the connected IP address are processed. Because log file analysis is not affected by any time-based difficulties known as the probe effect, log files are frequently the sole means to identify and locate an error in software. As a result of the increasing speed of data on the web, a framework is created to handle and analyze data for vulnerabilities.
The purpose of this research is to monitor real-time logs, system behavior, and odd activity across the stack, and to track issues back to their underlying cause by analyzing them in the context of the complete stack by monitoring critical resources.

LITERATURE REVIEW

Shrish Warma and Dilip Sisodia,a review of Web Usage Pattern Analysis Using Web Logs, the ninth international conference on computer science and engineering. The authors examined the process of identifying useful patterns in an academic institute's web server log file. The findings obtained can be used in a variety of applications such as web traffic analysis, efficient website administration, site updates, system improvement and customization, and business intelligence, among others.

Michal, Munk et al., "Pillar 3-Pre-Processed Web Server Log File Dataset of the Banking Institution."- The server logs include a wealth of helpful operational data and failure warnings. The problem is that such information grows with time as a result of the amount of entries in the logs, which can quickly become unmanageable. The analysis of server logs must be automated in order for the logs to be used as a proactive administrator tool. Log analyzer is software with a three-tier design that parses log files generated by any web server that adheres to common web server logging standards. Analyze parsed data and categorize it into useful reports that the user can consume for administration or monitoring.

Mahesh, Manchanda. "Web Usage Mining: A Dynamic Methodology to Preprocessing Web Logs."-This research introduced a new dynamic preprocess technique for creating a dynamic training dataset for a prediction model using web mining, as well as Graph based substructure Pattern Mining (GSPAN) for enhanced preprocessing with proxy log. The proposed model would help to reduce cache size by 40%, boosting overall performance.

Siva Jyothi, Barla et al., "Recommending Based on Analysis of Users' Behavior in an E-Commerce Website."- Web mining is a procedure that is important in analyzing the behavior of web users. Web usage mining, which is a subset of web mining has a significant impact on web personalization. The goal of this project is to extract relevant information from web access log files and use data mining techniques to analyze customers' online shopping behavior.

Extracting Siam University's weblog for understanding user behavior on MapReduce - Map Reduce is a framework that lets developers to create applications that process and analyze vast amounts of data in massively parallel fashion. Furthermore, a click stream is a record of a user's Internet activities. We can gather, analyze, and present aggregate data about which pages users view in what order - and which are the result of the succession of mouse clicks each visitor makes using a click stream analysis. Click stream analysis can show usage trends, which can lead to a better knowledge of user behavior. In this research, we present an innovative and effective web log mining methodology for clustering web users.

II. RESEARCH METHODOLOGY

Web server log analysis emphasises the significance by putting an algorithm which is relatively easy for implementation and it gives impressive response. Web server log analysis system is a type of application that will help to detect and avoid attacks that may occur on the server by using the server's log files. These log files can be used to accurately give us an idea of where and how the attack was initiated and is taking place. In our work, we are having seven modules such as Data collection, preprocessing, data cleaning, Session identification, User identification, Pattern analyzing and output.

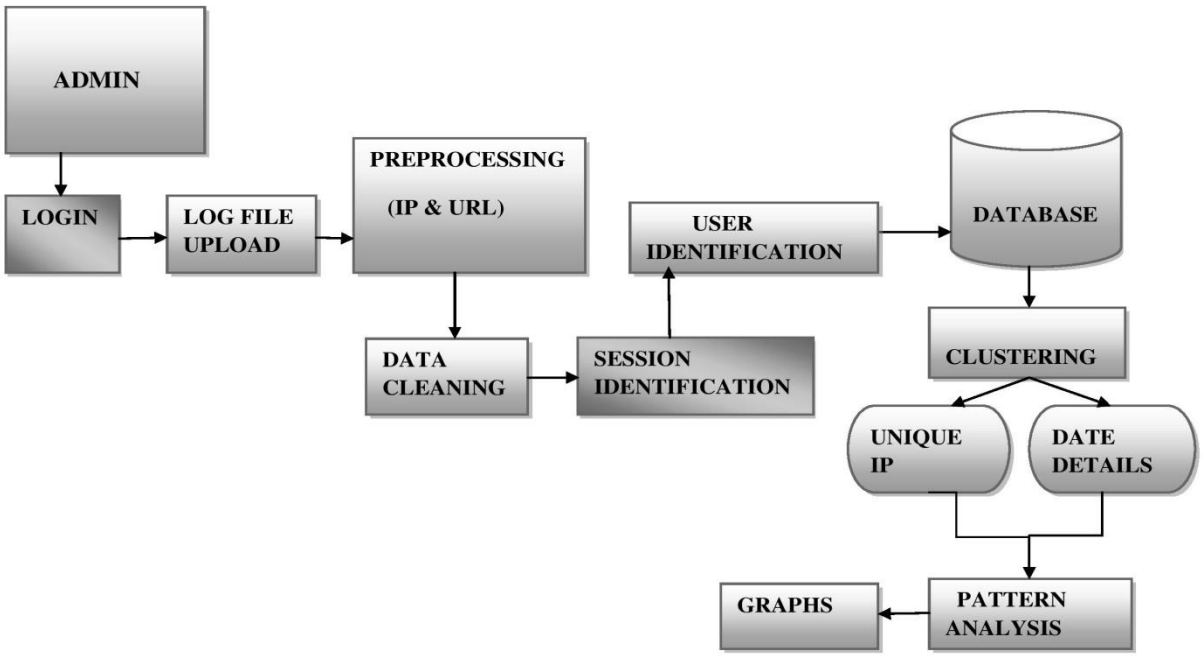

Fig.1 Architecture Diagram

In the above Figure 1, the admin upload the log data set then it can be preprocessed by cleaning the data and the log details can be stored to the database. The unique IP and date details log can be clustered using K-means algorithm. Finally the output can be in a graph format.

3.1 Data Collection

The generation of a sufficient preprocessed usage data collection is an important effort in web usage mining applications. In web usage mining, this technique is typically complex and critical to the successful extraction of usable information from log files.

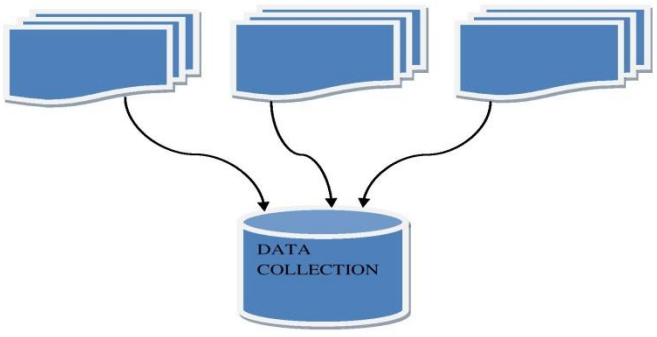

Fig.2 Data Collection

3.2 Preprocessing

Data preprocessing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. Preprocessing means cleaning the collected data to remove irrelevant information and preparing a log.

3.2.1 Pseudo code

```
words=[lemmatizer.lemmatize(w.lower())forwinwordsifwnotinignore_words]
words = sorted(list(set(words)))
classes = sorted(list(set(classes)))
print(len(documents), "documents")
print(len(classes),"classes",classes)
print(len(words),"uniquelemmatizedwords",words)
pickle.dump(words,open('texts.pkl','wb'))
pickle.dump(classes,open('labels.pkl','wb'))
```

3.3 Data Cleaning

The initial stage in the preprocessing of web usage Mining is data cleaning. Not all log entries in raw logs are appropriate for pattern analysis; we only want to save the records that contain important information. As a result, the data cleaning phase is utilized to remove extraneous items from the access log files.

3.4 Session Identification

A session is defined as the set of activities carried out by a user from the time he enters the website to the time he exits it. As a result, session identification is used in the process of segmenting each user's access log into individual access sessions.

Fig.3 Session Identification

3.5 Pattern Analyzing

For web usage mining, data mining techniques were employed to extract patterns of usage from web log files. Pattern discovery is a critical step in web mining that incorporates algorithms and approaches from a variety of study disciplines, including data mining, machine learning, statistics, and pattern recognition. To uncover rules and patterns, approaches such as statistical analysis, association rules clustering, classification, sequential pattern, and dependency in modeling are utilized.

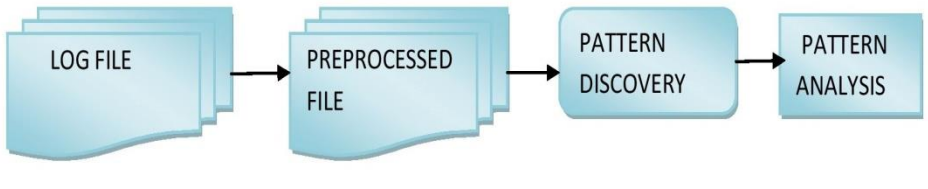

Fig.4 Pattern analyzing

3.5.1 Steps

The data set can be reads from the log fie.
The IP addresses can be extracted.
By using K-means algorithm the clustering begins by grouping of IP addresses.
It may use to taking out the unique IP addresses it may eliminate the duplicate entries of IP addresses.
The IP addresses are validated with the help of regular expression.
Then the graphs can be given as the output while it may have the IP wise graph and date wise graph according to the requests from the IP addresses to the server.

3.6 Output

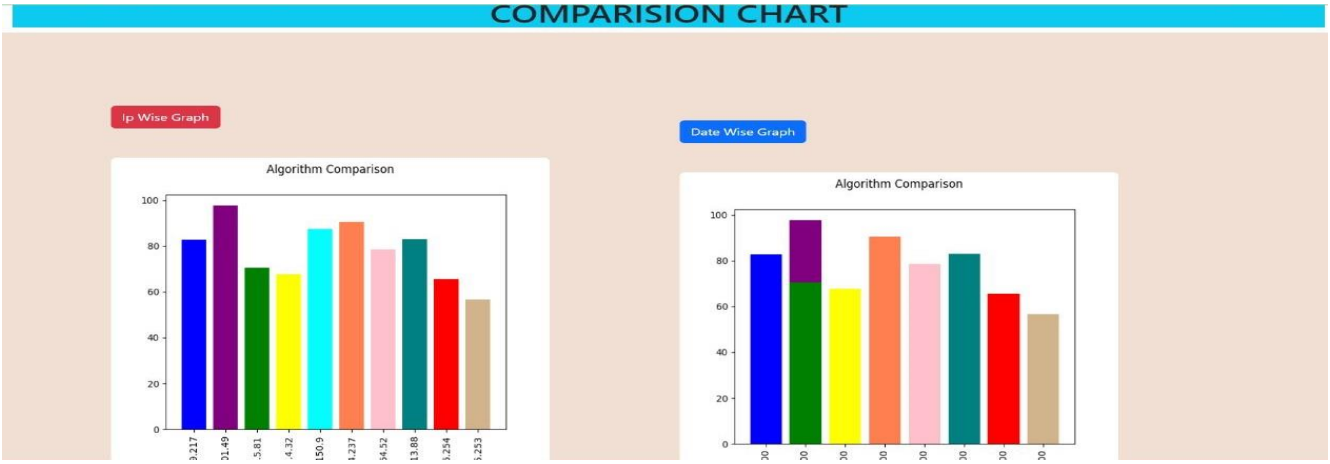

Fig.5 Sample Output

III. RESULTS AND DISCUSSION

The typical log process proposed the technology does preprocess- store and directly visualize, resulting in lower overall latency. Using the suggested platform, real-time Data Analytics on massive data sets is achievable, allowing for quick insights into the data. The platform can accept any sort of log file, giving it the flexibility to evaluate several logs at the same time. The platform can also be developed as a foundation module to service other applications for real-time analytics-based decision making. The clusters, which are generated on the fly based on the Patterns detected, aid in reducing search time. As a result, instead of exploring the entire data store, the search is performed in a single cluster. This significantly lowers the time between request and response. As an added feature, the platform can be modified to serve any type of data other than logs if the data meets the system's regular expression. As a result, the offered remedy will be implemented.

Data mining techniques like association, clustering, and classification can be applied only to the group of interested regular users to find frequently accessed patterns, which results in less time consumption and less memory utilization with high accuracy and performance. This is preferable to tracking the behavior of all users (interested or not interested) in order to redesign the website to support user.

4.1. Sample Screenshots

Figures 6 to 9 represents the sample screenshots.

Fig.6 Home Page

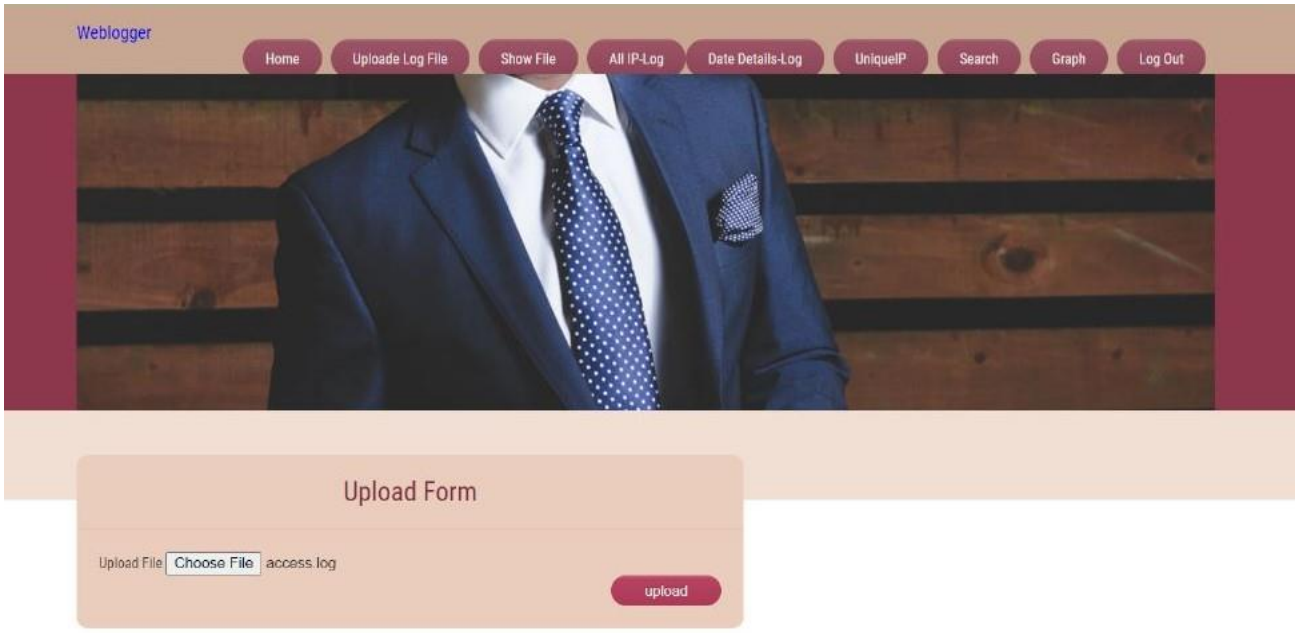

Fig.7 Upload log form

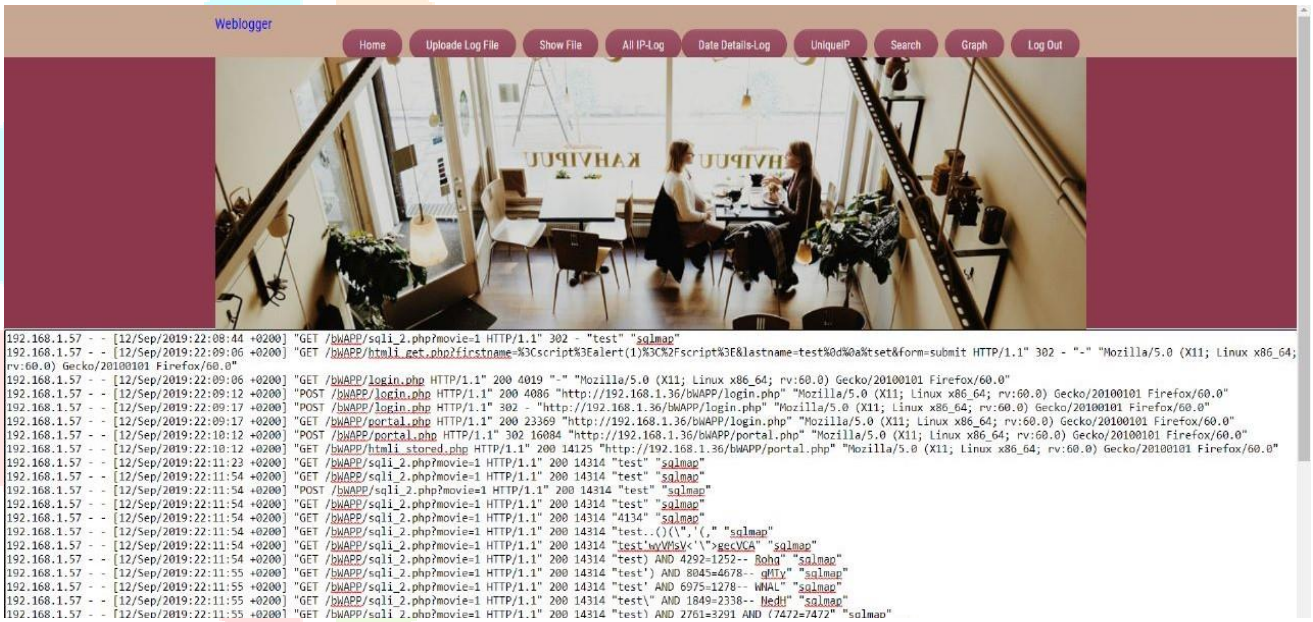

Fig.8 Data set

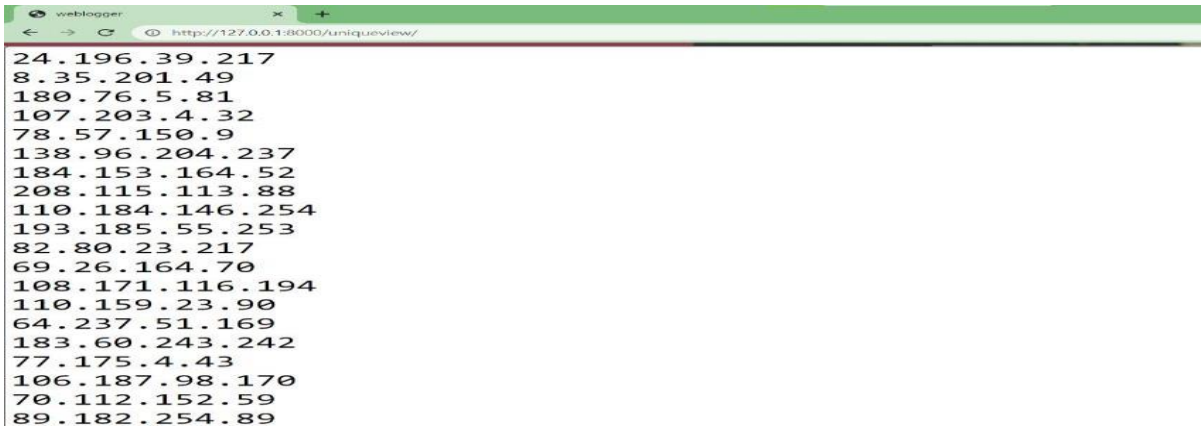

Fig.9 Unique IP

V. FUTURE ENHANCEMENT

In the web server log analysis, only the set of users who are interested in finding frequently visited patterns can employ clustering, which results in great accuracy and performance with low memory usage and time consumption. For future enhancement, it can read any kind of log files, giving it the ability to study several logs together. The platform can be expanded to support different applications for decision-making based on real-time analytics as the foundational concept.

VI. ACKNOWLEDGMENT

We would like to acknowledge our sincere thanks to the Management of our College and our family members who have supported and helped us in different stages of this project work.

REFERENCES

[1] Dilip sisodia and shrish warma Web Usage Pattern Analysis Through Web Logs: A Review 2012 ninth international conference on computer science and engineering.
[2] Munk, Michal, et al. "Pillar 3–Pre-Processed Web Server Log File Dataset of the Banking Institution." Data in Brief, vol. 39, Dec. 2021, p. 107672. DOI.org (Crossref), https://doi.org/10.1016/j.dib.2021.107672.
[3] Extracting weblog of Siam University for learning user behavior on MapReduce,2017
[4] Barla, Siva Jyothi, et al. "Recommending Based on Analysis of Users Behaviour in An E-Commerce Website." International Journal of Computer Sciences and Engineering, vol. 6, no. 5, May 2018, pp. 816–20. DOI.org (Crossref), https://doi.org/10.26438/ijcse/v6i5.816820.
[5] Manchanda, Mahesh. "Web Usage Mining: Dynamic Methodology to Preprocessing Web Logs." HELIX, vol. 8, no. 5, Aug. 2018, pp. 3810–15. DOI.org (Crossref), https://doi.org/10.29042/2018-3810-3815.