International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 52-58

Command Line Web Log Analysis through Open Source Utilities

G.Thangaraju[1], J.Umaranii[2], R.Ramkumar[3]
[123]*Research Scholar*
[1]*(Department of Computer Science, /Karpagam University,/Coimbatore)*
[2]*(Research and Development Centre, /Bharathiyar University,/ Coimbatore)*
[3]*(Department of Computer Science, Thanthai Hans Roever College (Autonomous) Elambalur, Perambalur)*

Abstract: *In the current scenario of communication technology, in the respect of the needs of individual, officials, or enterprise administrative and government sectors all of them are communicate with the internet, based on their services and needs. So, the all transmissions are done via the some of the organized technical aspects like the web pages, web server, on line services rendered by the different administrative or business peoples. The end user can acquire their needs and their responsibilities to reach their task with the help of internet technology. In this research focus pre-processing stages of the web usage mining technology with the help of the command line open source utilities to analysis the web log files. This work organized as follows; the first section consist the introduction to web mining and web usage mining, pre-processing stages of web usage mining and details of log file structure and types. The second section describes the related works. The third section describes the methodology and the fourth section describes the experimental results. The fifth section delivers the conclusion of this research work.*

Keywords: *Web usage mining, Pre-processing stage, Web log files, and Open source utilities.*

I. Introduction

The cosmic growth of the web technology the users also in the huge level, the technology covers the information's as in the following aspects; semi-structured, structured , unstructured, static, dynamic, video, audio, distributed and high dimensional data's available in the www in the form of web pages. The information's is available in large and easily accessed by the users, each and every moments of the web users are monitored by the web browser and its stored in the form log data and it is stored in some location of the client and server level.. [Valsamidiset.al., 01], The web page intact with server and client through the request and response technical aspects, and also maintain the both the server and clients communications are kept the log file in the respective locations.

Web mining is the one of the part of data mining applications; it is classified into three types based on the information to be mined, the classification depicted as in the following diagram

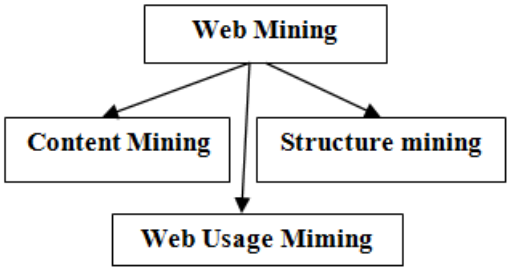

Fig 1. Classifications of Web Mining

1**. Content Mining:-** It is one of the web mining types, it is the process of extracting useful information from the web documents.

2. **Structure Mining:-**The structure of the Web can be represented as a network where the web pages are nodes and the hyperlinks that connect two related pages are edges between any two nodes.

3. **Web Usage Mining:-**Web Usage Mining is the one of the part of Web Mining, which is used to mine the web log data/information to discover useful patterns which can be exploited as user personalization and navigation, to gather the useful information like to identify the user frequently accessed the site with their location, some of the user occasionally use the particular site, how the web pages are interconnected with the access time like the user/analyzer can decide their questions [Sahaj Chavda et al,09]. Based on the user question the analysis will be done and produce the results.

A). Web Usage Mining

The web usage mining consists the four basic phases; they are 1. Data Collection 2. Data Pre-processing 3. Pattern discovery 4. Pattern Analysis.

1. Data Collection: -It is the first of the WUM it is to collect the data from the different resources among the web

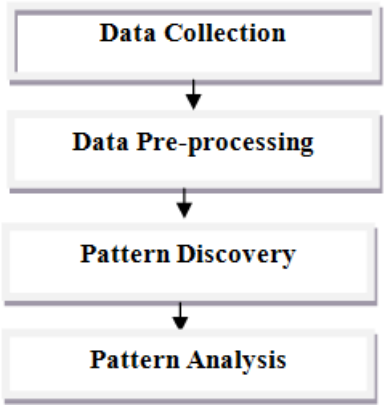

Fig .2: Four Phases of Web Usage Mining

Figure 2 represents the different phases of web usage mining; they are discussed in individually in a specified manner.

2. Data Pre-processing:-It is the next phase after the data collection; huge amount of data collected from the previous phase, the collected data should be consistent and integrated in order for them to be used in pattern discovery. It includes Data cleansing, User identification, Session identification and Path completion.

3. Pattern Discovery:-In this stage, the pre-processed data's is analysed to extract valuable patterns. Statistical methods and machine learning are used to mine patterns. The following approaches used for pattern discovery path analysis, association rules, clustering, classification, sequential patterns and order model discovery [R.Padmapriya et al, 10].

4. Pattern Analysis: -After the completion of pattern discovery, some of the techniques and tools are needed to make these patterns understandable for analysts and to maximize the benefits from these patterns. Techniques include database querying, graphics and visualization, statistics and usability analysis.

B).Pre-Processing Of Web Usage Mining

In the pre-processing stage of Web Usage Mining is consists the sub operations like Data Cleaning, User Identification, Session Identification and Path Completion. Each operations can have on own roles and responsibilities, they are summarized as below;

1. DataCleaning: -Data's collected via the data collection process, the collected data have a noisy or unwanted data's are kept with the log file. So that remove the noisy like blank spaces, repeated information's, audio, video, images are removed via the data Cleaning Process.

2. User Identification:-It is to identify which user accesses the webpage and website . Many methods have been proposed to automate user identification and the most well-known are: IP Address, Cookies and Include a unique ID generated by the web server in the URL instead of using the cookie file.

3. Session Identification:-Session identification, which encodes the navigational behaviour of the users, is very important in usage mining. A user session is a series of web pages that the user visits in a single website access. Various methods have been proposed to identify user sessions. They are as follows; Time-based methods, Specific time period, Context-based methods and examine the time spent on the page.

C).Web Log Files

Web Log file is a file automatically generated and monitored by a web server. Each and every Hit to the web site, including each view of an web page, like images, audio file and other types of the content in web page. The raw web log file is one line of text for each hit to the web site [Dafa-Alla et al, 08]. This consist the complete access details of a web page. In internet arena have numberof internet service provider and web servers are available, each web server can have their log file format, the Common Log File Format is also available. In this work focus on the Apacheweb server. The following are locations of web log files in Linux server with different flower;

1./var/log/httpd/access_log
2./var/log/apache2/access.log
3./var/log/httpd-access.log

The location and content of the access_log are controlled by the CustomLog directive. The syntax of CustomLog are as follows; CustomLogfile|pipeformat|nickname [env=[!]Environment-variable].

Apache web server maintains the log files mainly in two categories 1. Error Log 2. Access Log, in this research work focused on the Access Log only.

Apache web server Access Log consist the some type of log formats they are listed below;

1. Common Log Format
2. Combined Log Format
3. Multiple Access Log
4. Conditional Logging

The following sections give the essentials of the access log types;

1. Common Log Format

This the first type of access logs. The following are general Common Log Format
LogFormat "%h %l %u %t \"%r\"%>s %b" common
CustomLog logs/access_log common
The following line describes the example of the above CLF-Common Log Format;
165.0.0.1 – rctls [10/Dec/2017:14:53:25 -0700] "GET
/apache_pb.gif HTTP/1.0" 300 2547

Table.1: Common Log Format Descriptions

S.No	Field in the CLF	Descriptions
01	165.0.0.1 (%h)	IP Address
02	(%l)	Indicates the requested piece of information is not available.
03	rctls (%u)	User id
04	10/Dec/2017:14:53:25 -0700] (%u)	The time that the server finished processing the request(day,month,year,hour,minute, zone]
05	"GET /apache_pb.gif HTTP/1.0" (\% r\)	Request line from the client within the double quote
06	300 (%>s)	Status code
07	2547 (%b)	Size of the object returned to the client

2. Combined Log Format

It is second type access log, it is also commonly used log formats in apache web server
LogFormatLogFormat "%h %l %u %t \"%r\"%>s %b \ " %{Referrer} i\" \ "%{User-agent} i\"" combined
CustomLog log/access_log combined

Combined log file format have the fields as in the common log format upto the %b field, remaining things are additional options in the combined log file, http://www.bdu.ac.in/home.html(\"%Referer) i\"),
The "Referer" (sic) HTTP Request Header "Mozilla/5.0 [en](Windows; I; Nav"(/"%(User-agent)i\"
This gives the web site referred by the client. The UA-User Agent HTTP request header. This is the browser used by the client.

3. Multiple Access Log

It is third access log It can be created by the CustomLog directives in the configuration file. Consider the following example consists the three multiple access log. The first represents about the CLF, second represents the Referer and the third represents the browser information.
The last two CustomLog lines show how to mimic the effects of the RefererLog and AgentLog directives.

LogFormat "%h%l%u%t\"%t"%>s%b"common
CustomLog logs/access_log common
CustomLog log/refere_log "%{Referer}i->%U"
CustomLog logs/agent_log "%{User-agent}I"

4. Conditional Logging

It is the fourth type of access log. Based on the client request some conditions to apply the existing log, it gives the specific result based on the query. For that purpose one variable can be used, it is named as the SetEnvIf, assign the value to this variable. It is referred by the log and produced the respective values.

II. Related Works

The authors proposed a methodology to improve the content quality of Learning Management System courses with the application of new metrics and existing data mining clustering algorithms K-means and Markov Clustering Algorithms. The authors show that the proposed metrics can offer a preliminary course ranking, which in turn can be used as input for clustering algorithms. These suggest specific actions to instructions so that they can improve their content, course usability and help them adapt their courses to student capabilities [Valsamidis et al.,01].

The authors concludes that various applied data pre-processing techniques like user behaviour and navigations, total no.of users has visited the web page, which time they accessed and with their advantages and disadvantages are discussed[Mitali Srivastava,et al,02].

Pre-processing the server log files with the different steps through the SQL query and java code, remove the noise and unwanted records from the log file [Rachit Goel,03].

Pre-process the log files through the java source code for filed extraction, the collected log files are merged into one file through the improvised merging algorithm, then clean the log files with the removal of unwanted fields and finally convert the text file into table with the oracle table format. The task of the proposed work is to identify the sessions with the different heuristics. The authors done the mentioned process as in well [Sanjay BapuThakare et al,04].

The work focus on techniques and architectures for more effective integration and Mining of content, usage, and structure data from different sources is likely to lead to the next generation of more useful and more intelligent applications, and more sophisticated tools for Web usage mining that can derive intelligence from user transactions on the Web [K.S.R.Pavan Kumar et al.,05].

A paper [Navin Kumar Tyagi et al.,06], consists the two proposed algorithms one for data cleaning and another one for data reduction. The first one remove the files with the extension of .jpg,gif,.css and the status codes keep as it is. The status code removed via the second algorithm with the identification of incomplete and completed session entries.

A paper [Jia Li, et al., 07] explains the web log file cleaning process for mining the log data. Presented an overview of web usage mining and provides two algorithms one for data cleaning another one for field extraction, the extracted data converted into the structured forma and mining the information from that by the use user proposed algorithm.

III. Methodology

Command Line Web Log Analysis Through Open Source Utilities:-

Open Source Software Foundation introduces the many of the software's with the free of the cost and free access. Among them Apache Web Server is one of the product, to serve the internet user community, it's provide their services 69.00 per cent of the internet market. In this research proposed a methodology to analyse the web log files using the open source utilities provided by the open source software, they are combined into a form of shell scripts to attain a task. This process started with collection of the web log files from the Apache web server and also analyse them with the help of open source utilities, finally produce the reports based on the user requested task, the following are depicts the architecture of the proposed methodology and their successive operations;

Step.1: Web Log File: - First of all the researcher collects the web logs from the web server, proxy server and browsers. The collected log files are combined/merged with one file The collection of web log file have the syntax based on the server, here we consider the one format to collect the web log file from the apache web server.

Wget-https://s3.amazonaws.com/secureNinja/pythan/access-log

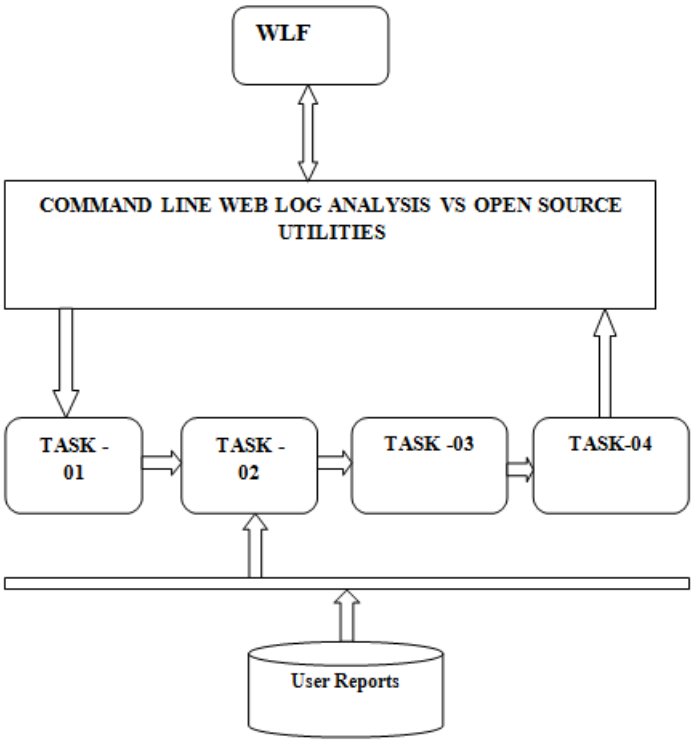

Fig.3: Architecture of the proposed system

Figure.3: represents the sequence of operations flow in this research.

Step.2: Command Line Web Log Analysis through Open Source Utilities (CLWATOSU)

Second step of the proposed architecture consists the five task to process the web log files, task accomplished by the user, based on the task the process will work and produce the report to the user.

Task 1: User can view the content of the log file in screen without editing.

Task 2: Find the text or any information as the need of the user.

Task 3: Find out the particular filed value only.

Task 4: Arrange the required line of tog with specific filed in ascending or descending order.

Task 5: Eliminates the duplicates, first/last part of the user requested part to display.

Mentioned above tasks are processed through the some of the open source utilities, the table 2 represents the open source shell scripts commands and their descriptions the task achieved through the shell scripts.

Table.2: Essential Open Source Utilities

S.No	Shell Commands	Descriptions
01	cat	Prints the content of a file in the terminal window
02	grep	Searches and filters based on patterns
03	awk	Can sort each row into fields and display only what is needed
04	sed	Performs find and replace functions
05	sort	Arranges output in an order
06	uniq	Compares adjacent lines and can report, filters or provide a count of duplicates.
07	head	First/starting part of the file
08	tail	Last/ending part of the file

The mentioned above shell commands are do their work with appropriated options in this work use the first five commands within the file, tasks are written with their syntax by the use of case control statements and necessary scripts commands are used

Algorithm: Command Line Web Log Analysis through Open Source Utilities (CLWATOSU)
Input: Combined_Access_log file
Output: Reports based on task
Step 1: Read Clog_file.txt
Step 2: Read T
Step 3: case $T in
Step 4: 1) cat Clog_file.txt ;;

```
Step 5: 2) ls | greplog ;;
Step 6: 3) awk –F\" ‘{print $6} ‘ Clog_file | sort |uniq –c| sort –fr>Clog_fileNew ;;
Step 7: 4) sort Clog_fileNew ;;
Step 8: 5) head -50 unique Clog_fileNew ;;
Step 9: *) Are you Finished your task? ;;
Step 10: esac
Step 11: end
Step 12 : //save the script and run using the sh command.
```

Figure.4: Algorithm for proposed work

Figure.4: represents the execution flow of the user specified task, after the execution the results are produced, it will discuss the later sections.

Step.4:- User Reports:- During the execution of the combined access log file, they ask the option ,if we are given the first option the command will process and display the content of the log file within the minimum time period. Suppose we gave an second option search the log as the part of a file name, related information’s are displayed. Like the third, fourth and fifth task. In case if we are given other than 1 to 5 it will print the give message and return to the shell script prompt.

IV. Expremental Results

The following images are taken from the screen shot of the task processing in the Linux operating system.

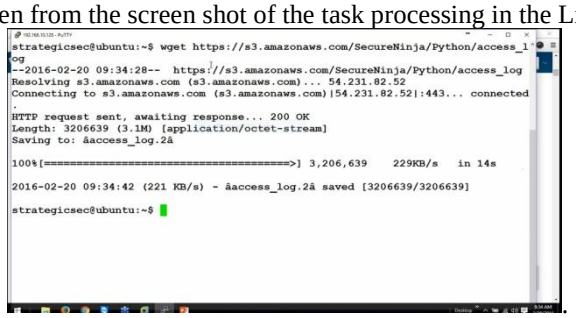

Fig.5: Data Collection command wget result

Fig .6: grep command processing

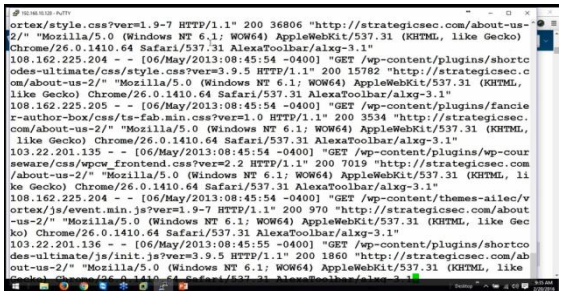

Fig.7: cat command processing

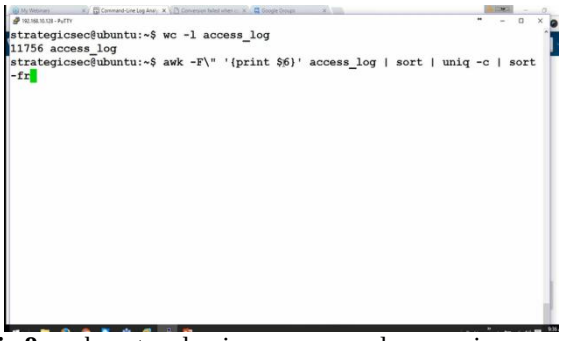

Fig.8: awk,sort and unique command processing screen

Fig.8: awk,sort and unique command processing output screen

The reports of the tasks are depicted in the figure 5, 6,7 and 8. Figure 5 represents the data collection and remaining are the task processing outputs.

V. Conclusion

Web log analysis is the impartment in the internet arena, researchers can research the web log files in their own view, like can collect the windows server log file and also windows browser log file analyze with the different predefined tools, in this paper represents complete structure to analyze the web logs with the Apache web server and open source utilities, this methodology give effective time consuming compared with the predefined tools.

References

[1]. Valsamidis,, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., &Karakos, A. (2012). A Clustering Methodology of Web Log Data for Learning Management Systems. Educational Technology & Society, 15 (2), 154–167.

[2]. Mitali Srivastava, RakhiGarg and P.K.Mishra, Pre-processing techniques in Web Usage Mining: A Survey, International journal of Computer Applications (0975-8887) Volume 97-No.18, July 2014.

[3]. Rachit Goel and Sandeep Jain, Improvisation in Web Mining Techniques by Scrubbing Log Files, International Journal of Advanced Research in Computer Science, Volume 5, No.5, May-June 2014.

[4]. Sanjay BapuThakare and Prof.Sangram Z Gawali, A Effective and Complete Pre-processing for Web Usage Mining, International Journal on Computer Science and Engineering Vol.02, No.03, 2010, 848-851.

[5]. K.S.R.Pavan Kumar, V.V.Sreedhar and L.ManojChowdary, A Critique on Web Usage Mining, International Journal of Computer Science and Information Technologies, Vol. 3 (5) , 2012,5276-5279.

[6]. Navin Kumar Tyagi, A.K. Solanki and Sanjay Tya , "An algorithmic approach for Pre-processing in web usage mining" International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 279-283.

[7]. Jia Li, "Research of Analysis of User Behavior Based on Web Log", 2013 International Conference on Computational and Information Sciences.

[8]. Dafa-Alla, Mirghani. A. Eltahir and Anour F.A, Extracting Knowledge from Web Server Logs Using Web Usage Mining, 2013 international conference on computing, electrical and electronic engineering .

[9]. Sahaj Chavda, Saurabh Jain, NikunjPanchal and Manisha Valera, Recent Trends and Novel Approaches in Web Usage Mining, International Research Journal of Engineering and Technology, Volume:04,Issue:04:April-2017,pp:1318-1321.

[10]. R.Padmapriya and Dr.D.Maheswari, A Novel Pre-processing method for web usage mining based on Hierarchal Clustering, International Research Journal of Engineering and Technology, Volume:04,Issue:04:April-2017,1517-1521