International Journal for Modern Trends in Science and Technology
Volume: 03, Issue No: 07, July 2017
ISSN: 2455-3778
http://www.ijmtst.com

# Web Server Log Analysis Using Web Usage Mining

Kavitha B[1] | Meyyappan T[2]

[1]M.Phil Scholar in Department of Computer Applications, Alagappa University, Karaikudi, Tamilnadu, India.
[2]Professor, Department of Computer Science, Alagappa University, Karaikudi, Tamilnadu, India.

**To Cite this Article**
Kavitha B and Meyyappan T, "Web Server Log Analysis Using Web Usage Mining", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 07, July2017, pp.-361-365.

## ABSTRACT

*Log Files are the data files which includes the log information's like the name of the user, Date and access time, request type, transferred bytes from server and from 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 [1]. The main theme of analyzing web usages with the help of log file helps to understand the users future needs and maintains the content of the website by based on the previous data. This paper discusses about the log files and management of log files. This paper also proposes a graphical representation of web usage results.*

***Keywords:*** *log file, IIS , web usage mining, web user*

## 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 a well-structured and has an internal storage to hold web data's [2]. Simultaneously the server maintains background files to display the pages to the client computers. When the request came from an IP address the server receives the request details from 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 Medias are converted into the hypertext and the client can get the entire media's as a response from server. The same procedure will be followed for every client by based on the request [3]. The log file is the data files which can hold all the transacted info from client to server and server to client. The log file resides in the server machine.

## II. PARTS OF LOG FILE

Log file parts as well as contents of the log file may differ and the contents are decided by the server. By basically the log file may contain the following information's:

**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. This 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 is the important one to know the access frequency. This field can maintain the date of access and the time. Every access will be maintained as 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 the server 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 by based on the availability of ports when a client is requesting pages or server data's.

**Bytes Sent:** The value of sent data's are 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 upload bytes. This field can maintain the transferred data's from client to server. The total received bytes are calculated and maintained under this title.

These are the basic and fundamental data's 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.

## III. Availability of log files

The web server log files are created automatically at the time of sending requests to the server from client. The log files are created in the following machines:

1. The Server Machine

2. The Client Machine

### 3.1 Server Machine:

The web log a file which is stored under the server machine is can maintain the activities of client machines which access the web server from the browser. The Fields are may be similar as per our previous discussions. Also the web proxy server can maintain the log file similar to the server machine.

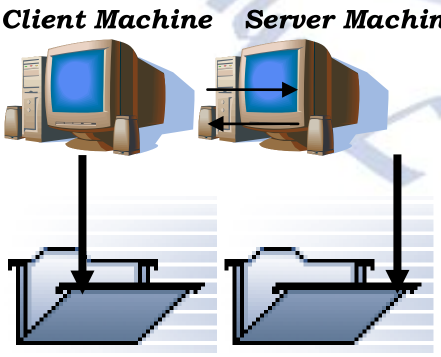
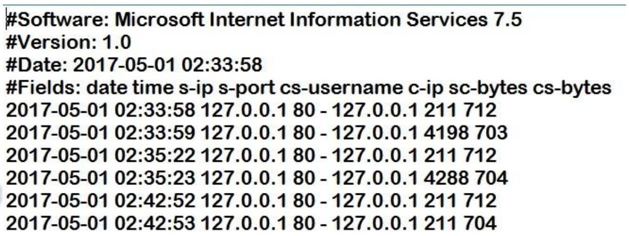


Fig: 1 Log Files

### 3.2 Client Machine:

The log files are also created on the side of client machines by itself. There are more number of software are available to generate the log files on the side of client. That kind of software can create and maintain the log file on the client machine. The servers have to decide to allow or deny the log file creation on the side of client machine.



```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2017-05-01 02:33:58
#Fields: date time s-ip s-port cs-username c-ip sc-bytes cs-bytes
2017-05-01 02:33:58 127.0.0.1 80 - 127.0.0.1 211 712
2017-05-01 02:33:59 127.0.0.1 80 - 127.0.0.1 4198 703
2017-05-01 02:35:22 127.0.0.1 80 - 127.0.0.1 211 712
2017-05-01 02:35:23 127.0.0.1 80 - 127.0.0.1 4288 704
2017-05-01 02:42:52 127.0.0.1 80 - 127.0.0.1 211 712
2017-05-01 02:42:53 127.0.0.1 80 - 127.0.0.1 211 704
```

Fig 2: Sample Log File

## IV. WEB USAGE MINING WITH LOG FILES:

Web mining is the task and process of mining and extraction of data's from the documents available in World Wide Web [4]. Web mining is a part of Data mining. The web mining is done by the various techniques to extract the data's from a large data set on the World Wide Web.

The web usage mining is described by the following figure:

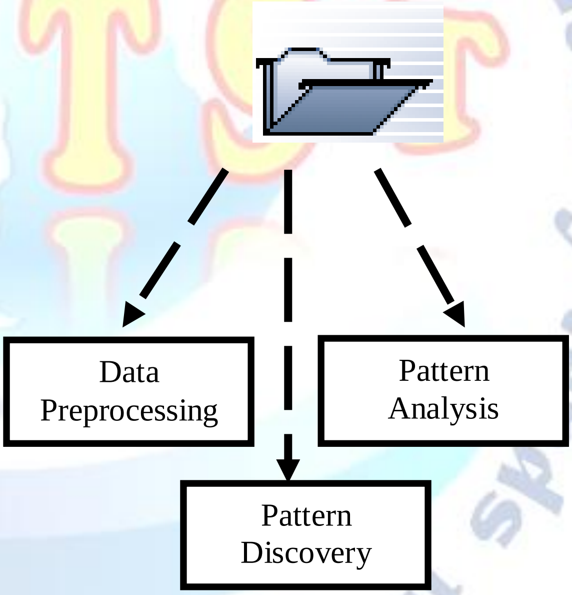


*Fig 3: Site Documents*

### 4.1 Data preprocessing:

The log file data's are called as raw data's and the raw data's are can't be utilized for data mining [5]. So it is necessary to convert the raw data's into readable data's. This conversion tasks are performed in the data preprocessing stage. These can be done in this research work by the following levels:

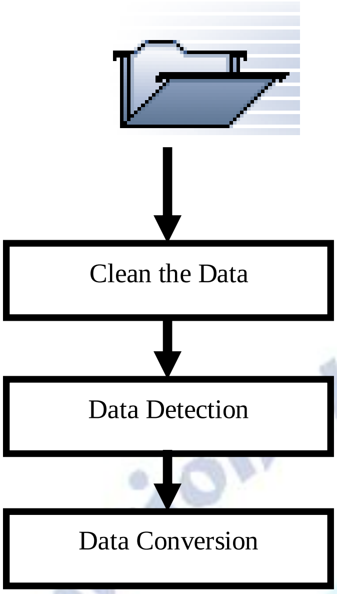
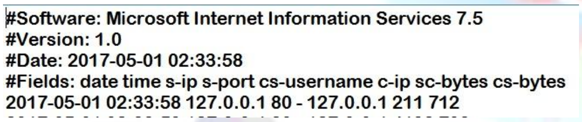


*Fig 4: Log File Processing*

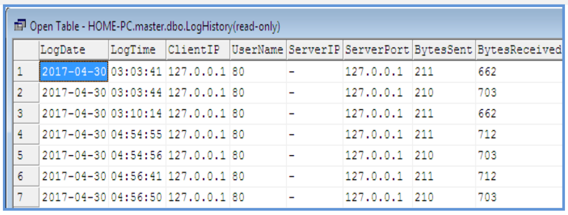
The *Data cleaning* stage is performed at first to clean the unwanted data records. For example, The Server log file creates the bookmark of the current server software and the version of the software. Also the log file shows the field headings in the top row. The header rows and the row data's are unnecessary. So, these kinds of data's are eliminated during the data cleaning process.



```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2017-05-01 02:33:58
#Fields: date time s-ip s-port cs-username c-ip sc-bytes cs-bytes
2017-05-01 02:33:58 127.0.0.1 80 - 127.0.0.1 211 712
```

*Data Detection:* The data detection is the second task performed in this research to discover the data's in the server log files. The data's are retrieved by using special mining based queries and the retrieved results are classified into categories [6]. This stage can collect the required data's like client IP, name of the client, sent and received bytes. The inbuilt code analyzes the data's 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 into the readable format of data's to perform mining methods [7]. In this research work the data's are converted and it will be maintained as a dataset with the help of SQL database. The sample converted format is shown below:



Open Table - HOME-PC.master.dbo.LogHistory(read-only)

| | LogDate | LogTime | ClientIP | UserName | ServerIP | ServerPort | BytesSent | BytesReceived |
|---|---|---|---|---|---|---|---|---|
| 1 | 2017-04-30 | 03:03:41 | 127.0.0.1 | 80 | - | 127.0.0.1 | 211 | 662 |
| 2 | 2017-04-30 | 03:03:44 | 127.0.0.1 | 80 | - | 127.0.0.1 | 210 | 703 |
| 3 | 2017-04-30 | 03:10:14 | 127.0.0.1 | 80 | - | 127.0.0.1 | 211 | 662 |
| 4 | 2017-04-30 | 04:54:55 | 127.0.0.1 | 80 | - | 127.0.0.1 | 211 | 712 |
| 5 | 2017-04-30 | 04:54:56 | 127.0.0.1 | 80 | - | 127.0.0.1 | 210 | 703 |
| 6 | 2017-04-30 | 04:56:41 | 127.0.0.1 | 80 | - | 127.0.0.1 | 211 | 712 |
| 7 | 2017-04-30 | 04:56:50 | 127.0.0.1 | 80 | - | 127.0.0.1 | 210 | 703 |

*Fig: 5 Converted Data's*

### 4.2 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 research work by the association rules [8]. The association rules are utilized in this research to find the correlation of data's 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's are grouped with the field headings.

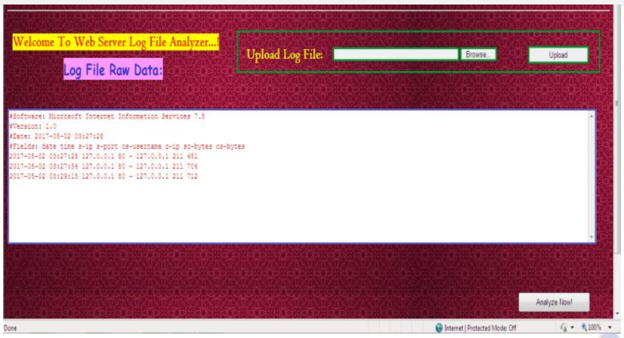
### 4.3 Pattern Analysis

The pattern analysis is the final stage on this research work. The analysis is the major task which can eliminate the irrelevant records from the required records. The pattern analysis is done in this research work by the help of Structured Query Language (SQL) [9].

The server administrator can get the converted data's 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 the server. The internal code eliminates the duplicates and the unique data's are shows 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.

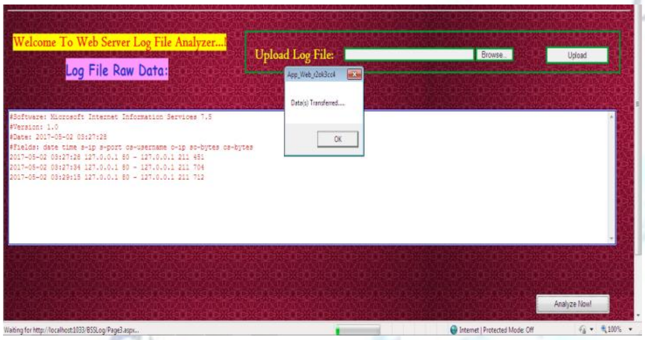
After the data's are associated into a single dataset the web page generates dynamic chart reports to show the results to the online user as well as the administrator of server. All these tasks and processes are performed internally by the proposed web application. The proposed research work does the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

The Various stages of results are shown in the following figures:


Welcome To Web Server Log File Analyzer...!
Log File Raw Data:
Upload Log File:
Browse..
Upload
Analyze Now!
Done
Internet | Protected Mode: Off
100%

```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2017-05-02 03:27:28
#Fields: date time s-ip s-port cs-username c-ip sc-bytes cs-bytes
2017-05-02 03:27:28 127.0.0.1 80 - 127.0.0.1 211 451
2017-05-02 03:27:34 127.0.0.1 80 - 127.0.0.1 211 704
2017-05-02 03:29:15 127.0.0.1 80 - 127.0.0.1 211 712
```

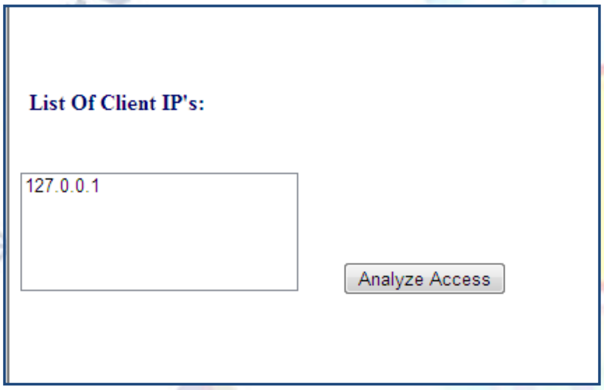
*Fig 6: Raw Log File*



Data(s) Transferred....

```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2017-05-02 03:27:28
#Fields: date time s-ip s-port cs-username c-ip sc-bytes cs-bytes
2017-05-02 03:27:28 127.0.0.1 80 - 127.0.0.1 211 451
2017-05-02 03:27:34 127.0.0.1 80 - 127.0.0.1 211 704
2017-05-02 03:29:15 127.0.0.1 80 - 127.0.0.1 211 712
```

*Fig 7: Data Conversion*



List Of Client IP's:

127.0.0.1

*Fig 8: Analyzed Pattern*

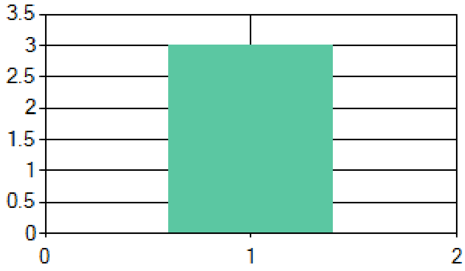


*Fig 9: Access Frequents*

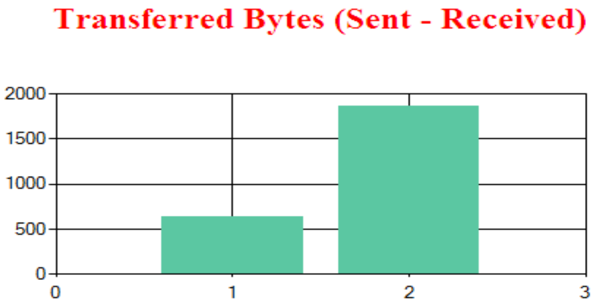


*Fig 10: Transferred Bytes*

## V. Conclusion

A web service can be more popular when it is offering solutions based on the expectation of users. 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 user's log. Analyzing log file is the easiest way to identify and understand the user needs.

This proposed 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 that for our needs.

## Acknowledgment

I would like to thank Dr.T.Meyyappan for his guidance and support for preparing this paper.

## References

[1] S.VijayalakshmiV.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 SystemsDesign and Implementation, San Francisco CA, Dec. 2004.

[4] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. Journal of the Royal Statistical Society B, pages 155–176, 1996

[5] J. Dean and S. Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters". Commune. ACM, 51(1):107–113, 2008.

[6] R. Cooley, B. Mobasher, and J. Srivastava, (1999) "Data Preparation for Mining World Wide Web Browsing Patterns," KNOWLEDGE AND INFORMATION SYSTEMS, vol. 1.

[7] Ratnesh Kumar Jain , Dr. R. S. Kasana1, Dr. Suresh Jain, (July 2009 )"Efficient Web Log Mining using Doubly Linked Tree," International Journal of Computer Science and Information Security, IJCSIS, vol. 3.

[8] K. R. Suneetha, and R. Krishnamoorthi, (April 2009)"Identifying User Behavior by Analyzing Web Server Access Log File," IJCSNS International Journal of Computer Science and Network Security, vol. 9.

[9] R.M.Suresh, R.Padmajavalli. "An Overview of Data Preprocessing in Data and Web Usage Mining". 2006IEEE

[10] BamshadMobasher et.al "Effective Personalization based on Association rule Discovery from Web usage data" WIDM01 3rd ACM workshop on Web Information and data management, November 9 2001, Atlanta 2001.

[11] Zhang Huiying, Laing Wei "An Intelligent Algorithm of Data Pre-processing in Web Usage Mining " Proceedings of the 5th world Congress on Intelligent Control and Automation, June15-19, 2004 Hangzhou, P.R.China.