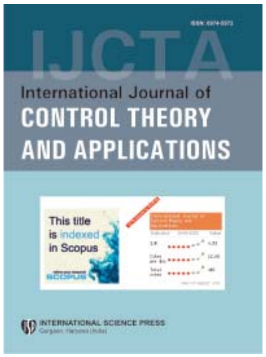

International Journal of Control Theory and Applications

ISSN : 0974–5572

© International Science Press

Volume 10 • Number 23 • 2017

Clustering of Web users Behavior based on the Session Identification Through Web Server Log File

S. Kalaivani[a] and K. Shyamala[b]

[a]*Research Scholar, PG and Research Department of Computer Science, Dr. Ambedkar Government Arts College (Autonomus), Affiliated to University of Madras, Chennai, India.*

E-mail: kalai5391@gmail.com

[b]*Associate Professor, PG and Research Department of Computer Science, Dr. Ambedkar Government Arts College (Autonomus), Affiliated to University of Madras, Chennai, India.*

E-mail: shyamalakannan2000@gmail.com

Abstract: Log file is maintained by web servers, which has the list of request made by the user. The content of the log files are required to be pre-processed within a significant amount of time. Even though we have many popular existing algorithms in data mining techniques, but still it's an open challenge for the researchers to improve them. In this paper, a system has been proposed, in which session identification is implemented based on the time spent on each page and also the number of web pages accessed by an individual user in a particular session is analyzed. Potential and non-potential web users were identified from the web server log file. Farthest first clustering algorithm and K-means clustering algorithm were used to identify the frequently accessed pages. These algorithms are implemented on the basis of weight factors such as time spent on the web page and maximum accessed page count. The experiment also shows that the efficiency of the farthest first clustering algorithm is better than K-means clustering algorithm for web log data. The experiment shows that the farthest first clustering has a significant improvement over K-means clustering algorithm in terms of the execution time.

Keywords: *Web usage mining, Pre-Processing, Session Identification, clustering, Web server log.*

1. INTRODUCTION

Web mining is a recent research area where a few need to improve the performance of the Website. Web mining is an application of data mining. It consists of three categories: they are content mining, structure mining, and usage mining. Structure mining and content mining make use of primary data on the web.

Web mining can be defined as the discovery and analysis of useful information from the Web World Web. With the explosive growth of information source available on the World Wide Web, it has become more important to find the valuable information from these huge amounts of data. Organizing websites with the useful content to the web users is not an easy task for the web designers. The solution of these problems can be provided by path analysis using web user navigation pattern.

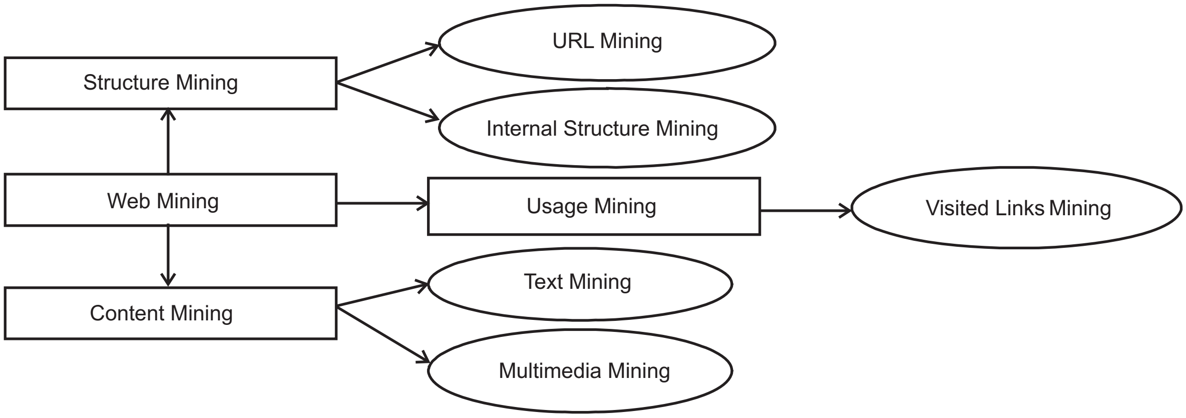

Figure 1: Classifications of Web Mining

Web user interest changes over the time but the static website will not satisfy user interest. The static website will become outdated and less attractive. The main objective of the website is to meet the user needs. In this paper an attempt is made to build an adaptive website which improves their user navigation based on the user access pattern. Figure 1 shows the classifications of web mining. User access patterns can be identified based on frequently accessed web pages from the web server log.

Web usage mining is one among the types of web mining. Web usage mining is used to identify an interesting pattern from web server log file [9], [10]. Web log is an interaction between the user and the website that automatically stored in the web server. Usage mining employs user browsing records to analyze user intentions. Web user has different purpose and intention while browsing a website. Each user's target and purpose can be predictable by an inspection of their browsing activities. Web usage mining involves three phases which is depicted in Figure 2 Pre-Processing is the initial step in the web usage mining. Pre-Processing includes Data Cleaning, Session Identification, User identification and Path completion [11], [12].

Pre-Processing
Pattern Discovery
Pattern Analysis

Figure 2: Phases of Web Usage Mining

1.1. Data Pre-Processing

Data cleaning technique is mainly focused to find unrelated, unpredictability, noise data and also to improve the quality of data [13]. Web server log file contains raw data and it is significant to extract the attributes from the file to eliminate inconsistent data. Usually log file data are separated using (,) or (“”). Field extraction plays a vital role in Pre-Processing where the data will be extracted from different fields. The main objective of Web usage mining is to improve the effectiveness of the websites by providing novel methods [14].

1.1.1. Data Cleaning

Web log pre-treatment work includes Data Cleaning, Session Identification, User Identification and Path Completion as shown in Figure 3 Data Cleaning is one among the process in Data Pre-Processing to remove unwanted data in the log file. Eliminating items which are irrelevant can be done by checking the suffix of the URL name. For example remove log entries with file name suffixes such as gif, jpeg, jpg. Web robots are also known as web spiders. It is a software tool which scans a website to extract its content. Error request is irrelevant for the mining process. If the status code is less than 200 and greater than 299 it must be removed.

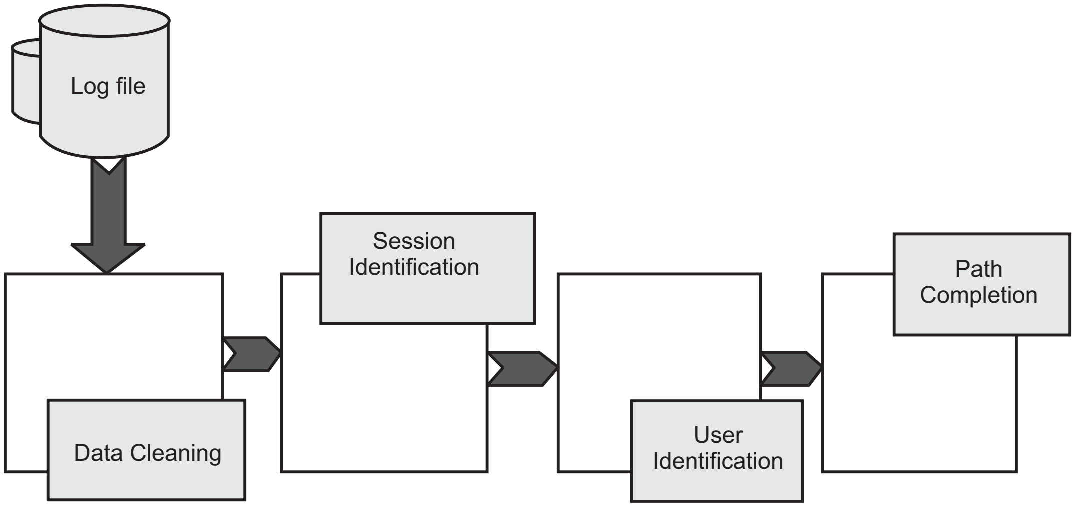

Figure 3: Data Pre-Processing Topology

1.1.2. User Identification

User identification consists of various methods to find a unique user from the log file. Various methods are discussed in Table 1. This work has identified a unique user based on the IP address. Different IP addresses represents dissimilar user and user identification is simplified.

Table 1
Methods of User Identification

No.	*User Identification*	*Method*
1.	User Identification	IP Address
2.	User Identification	Authentication Data
3.	User Identification	Cookies
4.	User Identification	Site Topology
5.	User Identification	Client Information

1.1.3. Session Identification

Once the unique user is identified, then it is important to find their session from their log file [15]. Session is nothing but set of request done by the identified user in a particular period of time. When the time gap between two repeated requests by the same user exceeds the certain threshold, then a new session is created.

$$\text{Session Creation} = s.t_{n+1} - s.t_n$$

Where $s.t_{n+1}$ and $s.t_n$ are the time stamp of two successive requests. Session of each user should be greater than 30 minutes. When the time spent on the web page by the user exceeds the threshold value, then only those pages are considered and then that web pages are taken to cluster.

1.2.4. Path Completion

The imperfect access path of every user session is predictable based on user session identification.

1.2. Pattern Discovery

Pattern Discovery is used to mine interesting patterns from log files by performing some data mining techniques such as Association rule mining, Clustering, and Classification.

1.3. Pattern Analysis

These patterns are stored and analyzed in the Pattern Analysis phase. It is the final step in the web usage mining. The purpose of Pattern Analysis is to sort out uninteresting information and to visualize only the interesting patterns to the user.

2. RELATED WORK

This section made a study of previous works which are related to the web log file usage and the approach used in analyzing them. World Wide Web has become a huge storehouse of data and serves as an important platform for the spreading of information. The users' accesses to Web sites are stored in Web server logs. Different works of several authors are discussed below:

The paper [1] proposed an approach to Pre-Process the Web log data and they have given a detail summary about the content of Web server log data. Each Uniform Resource Locator (URL) in the Web log data is separated into tokens and then it is implemented using SQL server management studio. Web server log data size is reduced by removing irrelevant files.

Another work reported in the paper [2] is concerned with the study of web server log data which has been collected from NASA website. Through analyzing the web log data using in- depth analysis approach, important information can be recognized such us top most errors and potential visitors. Future work of this paper is to use some of the data mining techniques to identify frequently accessed web pages.

Furthermore, Author [3] takes few mandatory information from the log file such as user-id, requested a web page, user session information and its content related meta-data for analysis from web server log file. The fusion of two analyses such us statistical and graph and graph analysis is done in this paper. Instead, the frequency count of a web page accessed the weight component time spent on the page is used in this paper.

In addition to the above, mishra [4] has used FP-Growth to explore the patterns which has been frequently accessed by the web user. FP-Growth is one of the suitable frequency mining algorithms which have extracted short frequent pattern as well as long frequent pattern. Apriori [5] is used in this paper for mining frequent pattern from the Web server log file database. The existing algorithms have been used by many researchers and it is extended by other researchers in the papers [6]-[8].

After looking at the existing algorithms, there are some drawbacks are found especially in terms of their performance. These drawbacks are mainly due to the inefficient method of scrutinizing an item or content of the log file. Typically string comparability will be used. If the log file contains multiple items, each of the items will be extracted first. After that, each number of transactions and the possible orders will be constructed.

The above process will need numerous scanning of web server log file which evidently incurs additional time cost. Moreover, the cost in terms of space usage is also vast when performing the analysis process. Hence, proposed approach is presented in section 5 takes care both of the issues mentioned above.

3. MATERIALS AND METHODS

Before presenting the log file analysis technique, this work is actually a little part of our superior research work which minimizes the log file size and cluster frequently accessed the pages with weight factor time spent on the page.

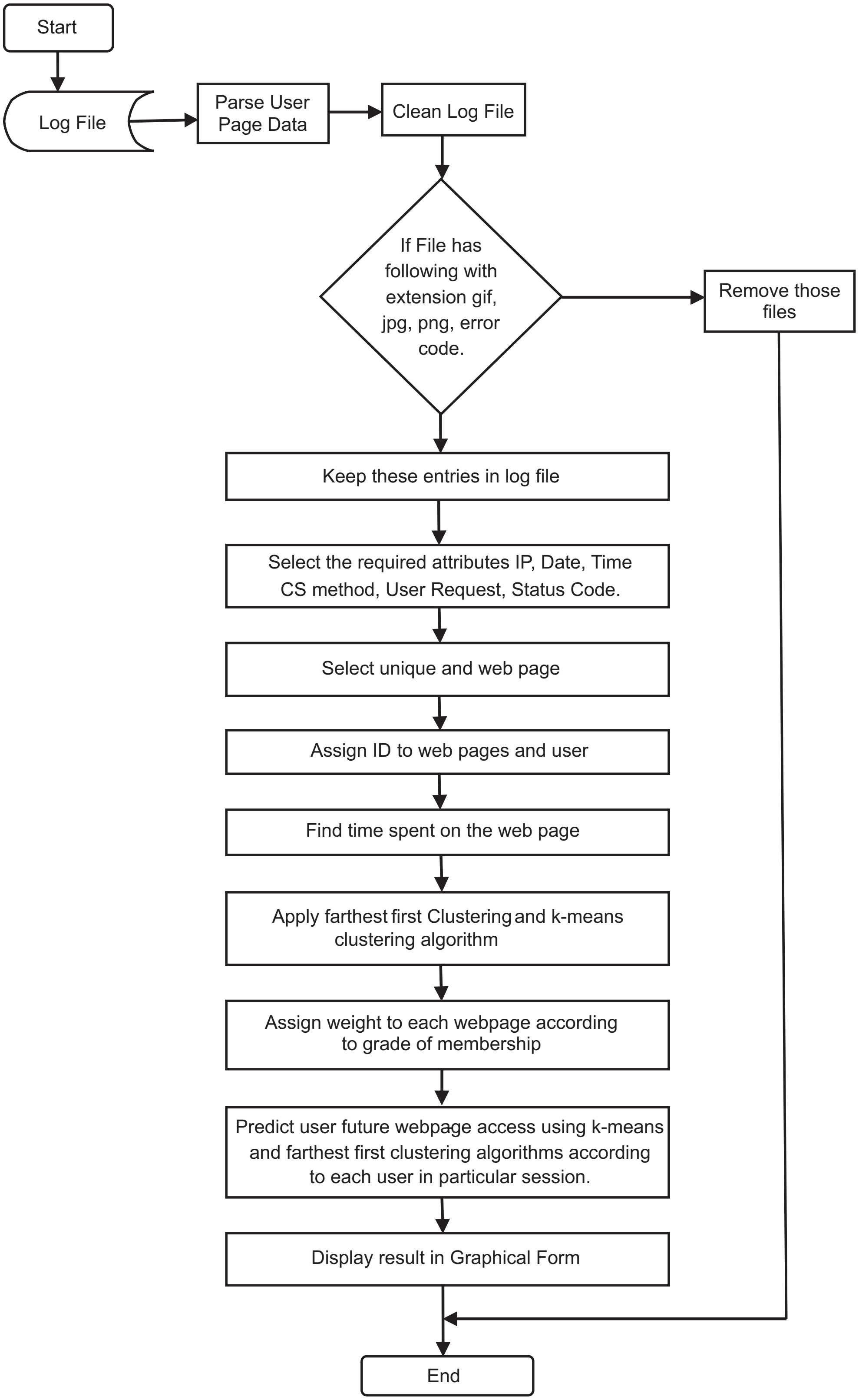
Start
Log File
Parse User Page Data
Clean Log File
If File has following with extension gif, jpg, png, error code.
Remove those files
Keep these entries in log file
Select the required attributes IP, Date, Time CS method, User Request, Status Code.
Select unique and web page
Assign ID to web pages and user
Find time spent on the web page
Apply farthest first Clustering and k-means clustering algorithm
Assign weight to each webpage according to grade of membership
Predict user future webpage access using k-means and farthest first clustering algorithms according to each user in particular session.
Display result in Graphical Form
End

Figure 4: Proposed Model of the System

Figure 4 depicts the proposed model designed for the study. As illustrated, the system model based on the web server log which was collected from NASA website [16]. Collected raw web log data contains irrelevant data. So, it should be pre-processed and then clustering is done to identify frequently accessed Web Page.

Algorithm for Data Cleaning

```
Input: Raw Web log Data
Output: Pre-Processed Web log Data
Begin
Step 1: Read Web log data from log file
Step 2: Parse web log data into tokens
Step 3: If (web log data! = null)
        Assign group based on the following Categories
Step 4: If (Web log data.url = ”*.jpg,*.gif, *.png*”) Then
        Assign as Unwanted Images Files
Step 5: Else if (Web log data.url =” *.flv,*. Avi, *. Mpg, *.mpeg”)
        Assign as Unwanted Video Files
Step 7: End if
Step 8: End if
Step 9: Save Records
Step 10: Repeat until the last record
Step 11: End
```

Algorithm for User Identification

```
Input: Pre-Processed Web log data
Output: Unique Users
Step 1: Read the Pre-Processed Web log data
Step 2: Compare IP address of two consecutive records
Step 3: If (IP address = same) then
Step 4: Assign as the Same User
Step 5: Else
Step 6: Distinct User
Step 7: End if
```

K-means Clustering Algorithm

In data mining, *k*-means clustering is a method of cluster analysis which aims to separate n observations into k clusters in which each assessment belongs to the cluster with the nearest mean value.

Steps to be followed

```
Step 1: Choose k objects from R (Records) as the initial cluster centers;
Step 2: repeat
Step 3: (re)assign each object to the cluster to which the object is closely similar, based on the mean value.
Step 4: Update the cluster mean value.
Step 5: Until remains no change
Step 6: End
```

Farthest First Clustering Algorithm

The farthest first algorithm proposed by Hochbaum 1985. It contains same procedure as *k*-Means clustering algorithm. Steps of the algorithm have mentioned below.

Steps to be followed

Step 1: Choose a random data as the center point first.

Step 2: Find the data which is the farthest point from the first point.

Step 3: Find the third point which is the farthest from the already existing point.

Step 4: Henceforth $i = 3,4,\ldots, n$ Find the data that has not been selected.

Step 5: It is the furthest point from $\{1,2,\ldots, i\text{-}1\}$ and mark it as point i.

Step 6: Use $d(x, \mathrm{S}) = \mathrm{miny}\epsilon\mathrm{S}\ d(x, y)$ to identify the distance.

Step 7: End

4. EXPERIMENTAL RESULTS

In this research work, set of experiments are done to evaluate the effectiveness and accuracy based on the log file prediction methods. To validate the farthest first clustering algorithm, we compared the result with famous *k*-means clustering algorithm. It is implemented in WEKA tool and the Pre-Processing algorithms are implemented in Ms SQL Server. The dataset used in this experiment is authentic, which is collected from the NASA website [16]. The dataset contains one-month records. In this procedure, first, the access log with the text format is loaded in the database. Then we need to remove the entire suffixes like **.jpg*,**.css*,**.gif* etc. These suffixes are not required in the file. Data Cleaning is done using Ms SQL Server.

Proposed Algorithm for Session Identification and to find Time Spent on Webpage.

Input: User Identified Web log data

Output: Maximum time spent on the web page and the potential web user

```
Begin
        Let P← Potential User;
        Let NP ← Non Potential User;
        Let T← Threshold Value(30 min);
        Let Pi← first entry of log file in a partic
        ular session;
        Let Pj← second entry of log file in same session;
    Compare two consecutive entries within the session;
        if timestamp(Pi) is less than timestamp(Pj) then
            The difference between the timestamp Pi&Pj should be greater than T;

        end
        If each user accessed more than five pages within the session then
            Assign P;
            Else
            Assign NP;
        end
End
```

The file size is reduced after cleaning the data. First image files, multimedia files, and incomplete URL are also removed using SQL query. Figure 5 shows the different kinds of analysis in the log file. Table 2 shows the summary of Pre-Processed data.

Table 2
Summary of Pre-Processed Data

No.	*Records*	*Count*
1.	Records in Original log file	16,37,264
2.	Records in Cleaned log file	5,88,830
3.	Attributes in Original log file	11
4.	Attributes in Cleaned log file	5

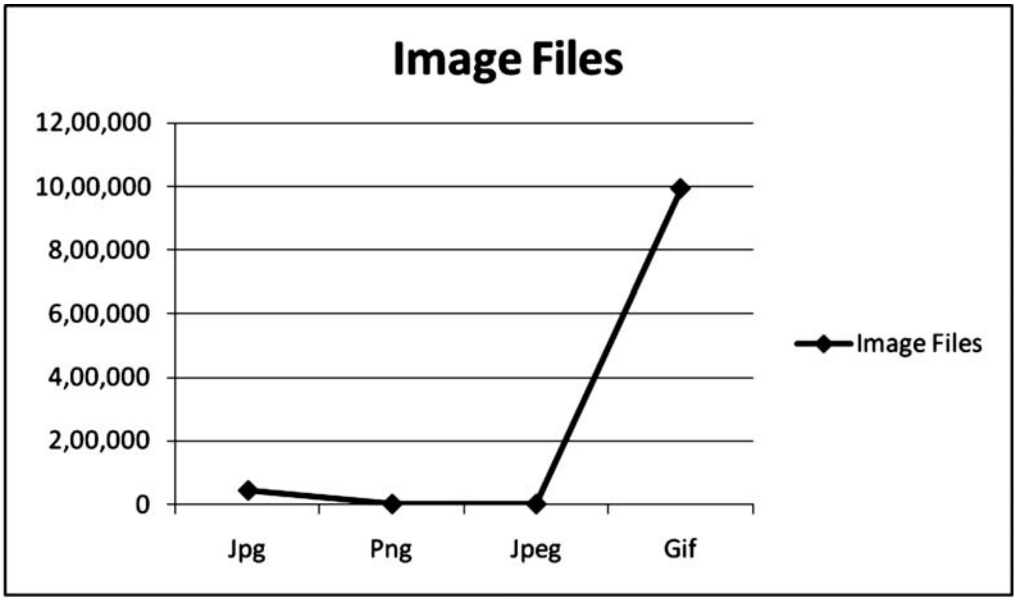

Figure 5: Image Files

Time oriented heuristics approach are based on the page stay time or time limitations. Time oriented approach is not optimal because they do not consider the web page connectivity. In navigation oriented heuristic approach artificially inserted backward browser movement is the major problem. Backward movement represents the reverse directions of edges. Another drawback on navigation oriented heuristic approach is the length of the session because it is used without any time limit. In the proposed algorithm, longer request sequences are constructed by using overall session duration time and page stay time. A sub-session can be created by using these two constraints. Only about 30% accessed 5 pages and less than 10% accessed 10 pages and more.

Running *k*-means clustering algorithm and farthest first clustering algorithm for the same dataset reveal that farthest first managed to outperform *k*- means in any data size of the transaction. Figure 6 shows frequently accessed web pages. Furthermore, farthest first clustering algorithm shows an incredible difference in time performance when the size of the dataset grows larger. Table 3 shows the result enhancement performance of farthest first clustering algorithm when compared with *k*-means. Figure 7 shows the results revealed the enhanced performance. The enhancement is in the sense of execution time taken to complete the entire process from reading the content of the log file to the finding of the most frequently accessed web pages.

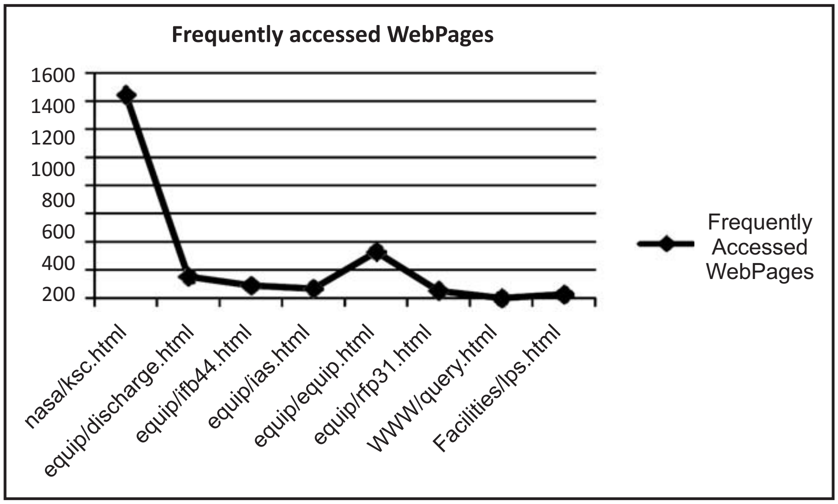

Figure 6: Frequently Accessed Web pages

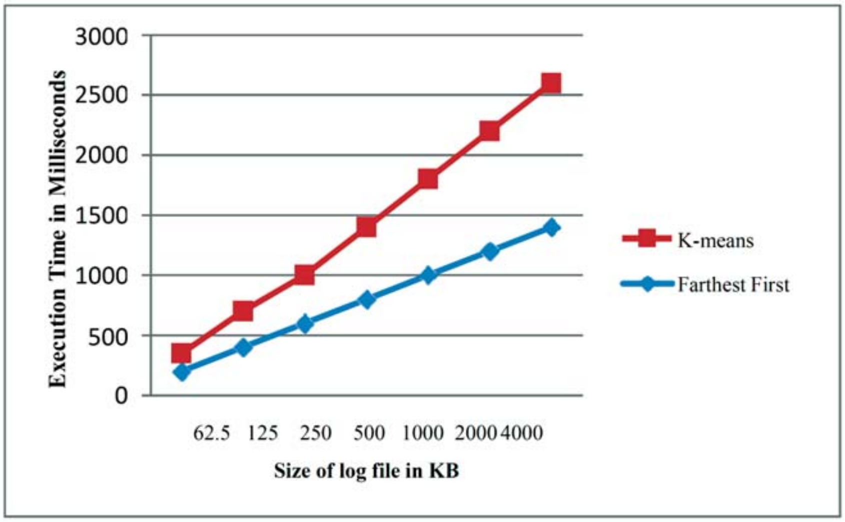

Figure 7: Execution Time of K-means vs Farthest First Clustering Algorithm

Table 3
Execution time of Clustering Algorithms

Size of the log file in KB	*Time taken to cluster*	
	k-means clustering algorithm(milliseconds)	*Farthest first clustering algorithm(milliseconds)*
62	200	150
125	400	300
250	600	400
500	800	600
1000	1000	800
2000	1200	1000
4000	1400	1200

5. CONCLUSION

Log files often contain enormous data which require a significant amount of time to be processed. Session identification is done based on the time spent on each page and number of web pages accessed by an individual user in a particular session. Two clustering techniques are performed on the Pre-Processed log data. Clustering is done to find frequently accessed web pages from the log file. When examining k-means clustering algorithm with farthest first clustering algorithm shows significant improvement in execution time performance. By implementing these techniques, faster analysis of the log file can be done even though the amount of data might be high. Reorganizing website based on the frequently accessed WebPages using data structure algorithm is considered as a future direction of research. Reorganizing website is mainly to reduce page access delay and to provide desired information in a fewer click.

REFERENCES

[1] S.Kalaivani, Dr.K.Shyamala, "A Novel Technique to Pre-Process Web Log Data Using SQL Server Management Studio," International Journal of Advanced Engineering, Management and Science. Vol 2(7). pp. 2454-1311, 2016.

[2] Suneetha.K.R, "Identifying User Behavior by Analyzing Web Server Access Log File," International Journal of Computer Science and Network Security. Vol 9(4). pp. 327-332, 2009.

[3] Stermsek.G, Strembeck, and M.Neumann, "A User Profile Derivation Approach based on Log file Analysis," In Proceeding of IKE. pp. 258-264, 2007.

[4] Mishra, R, "Discovery of Frequent Patterns from web log data by using FP - Growth algorithm for Web Usage Mining," International Journal of Advanced Research in Computer Science and Software Engineering. Vol 2(9). pp. 311-318, 2012.

[5] NehaGoel, C.K.Jha, "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool," International Journal of Computer Applications. Vol 62(2). pp. 0975 – 8887, 2013.

[6] Li,N, Li Zeng, Qing He, Zhongzhi Shi, "Parallel Implementation Of Apriori Algorithm Based On Mapreduce," 13th ACIS International Conference on Software Engineering Artificial Intelligence, Networking and parallel/distributed Computing. Vol 1(2). pp. 236-241, 2013.

[7] Jaishree Singh, Hari Ram, and J.S. Sodhi, "Improving Efficiency of Apriori Algorithm Using Transaction reduction," International journal of scientific and research publications. Vol 3(1). pp. 1-4, 2013.

[8] Ilayaraja.M, T. Meyyappan, "Mining Medical data to identify Frequent Diseases using Apriori Algorithm," Proceeding of the International Conference on Pattern Recognition, Informatics and Mobile Engineering. pp.194-199, 2013

[9] R.Suguna, D.Sharmila, "An Overview of Web Usage Mining," International Journal of Computer Applications. Vol 39(13). pp. 11-13, 2012.

[10] C.U.Om Kumar, P. Bhargavi, "Analysis of Web Server Log by Web Usage Mining for Extracting Users Patterns," International Journal of Computer Science Engineering and Information Technology Research. Vol 3(2). pp. 123-136, 2013.

[11] NehaGoel, Sonia Gupta, and C.K. Jha, "Analyzing Web Logs of an Astrological Website Using Key Influencers," International Research Journal. Vol 5(1). pp. 2-11, 2015.

[12] NehaGoel, C.K.Jha, "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool," International Journal of Computer Applications. Vol 62 (2). pp. 29-33, 2013.

[13] ShailyLanghnoja, MehulBarot, and Darshak Mehta, "Pre-Processing: Procedure on Web Log File For Web Usage Mining," International Journal of Emerging Technology and Advanced Engineering. Vol 2 (12). pp. 419-423, 2012.

[14] Latheefa.V, Rohini.V, "Web Mining Patterns Discovery and Analysis Using Custom-Built Apriori Algorithm," International Journal of Engineering Inventions. Vol 2(5). pp. 16-21, 2013.

[15] Catledge.L, Pitkow.J, "Characterizing browsing behaviours on the World Wide Web," Computer Networks and ISDN Systems. Vol 27(6). pp. 1-8, 1995.

[16] http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html.