Journal of Computer Science 9 (6): 771-779, 2013
ISSN: 1549-3636

doi:10.3844/jcssp.2013.771.779 Published Online 9 (6) 2013 (http://www.thescipub.com/jcs.toc)

Fast Real Time Analysis of Web Server Massive Log Files using an Improved Web Mining Architecture

[1]Ramesh Rajamanickam and [2]C. Kavitha

[1]Department of MCA, Sona College of Technology, Salem,TamilNadu, India
[2]Department of Computer Science, Govt. Arts College, Karur TamilNadu, India

Received 2013-03-18, Revised 2013-04-29; Accepted 2013-06-06

ABSTRACT

The web has played a vital role to detect the information and finding the reasons to organize a system. As the web sites were increased, the web log files also increased based on the web searching. Our challenge and the task are to reduce the log files and classify the best results to reach the task which we used. Aimed to overcome the deficiency of abundant data to web mining, the study proposed a path extraction using Euclidean Distance based algorithm with a sequential pattern clustering mining algorithm. First, we construct the Relational Information System using original data sets. Second, we here cluster the data by the Sequential Pattern Clustering Method for the data sets which make use of the data to produce Core of Information System. Web mining core data is the most important and necessary information which cannot reduce an original Information System. So it can get the same effect as original data sets to data analysis and can construct classification modeling using it. Third, we here used Sequential pattern clustering method with the help of Path Extraction. The experiment shows that the proposed algorithm can get high efficiency and avoid the abundant data in follow-up data processing.

Keywords: Path Completion, Cleanup the Data, Data Preprocessor, Travel Path Extraction, Sequential Pattern Clustering Method

1. INTRODUCTION

To discover knowledge based databases we commonly used one of the mining processes as Web data mining (Satish and Patil, 2011). We can implement Web data mining with the help of clear algorithms, software and many tools with the collection of information and analyze them from a large data set (Makkar *et al*., 2012). We use data mining process to extract only the elevated data, for analyzing the business.

We use the concept Web data mining process to extract the data from the World Wide Web. With effective and wide range of data we can make use of the internet with clear analysis of the user. Anyway by, retrieving and as well as filter the large databases is not an easy work. So that there are many web data mining tools are implemented to make extraction much easier. With the help of these tools we can detect the data and also understand the user needs.

Nowadays business uses have been increased largely to understand which we can expect precious information like summary of customers, analysis of the industry, business techniques using in corporate company. At any rate of time, the usage of this web mining is increased manifold.

In the busy world, data sets used by company has increased largely, so with the help of data mining, we can implement this tool to marketing, Customer Relationship Management (CRM), hospitals, scientific research, communications, financial services and other utilities. Some of the data are mining the test with the help of text mining, web mining, audio and video compressing data mining, network data mining, to find the relations in the large dataset (Shrivastava *et al*.,

Corresponding Author: Ramesh Rajamanickam, Department of MCA, Sona College of Technology, Salem,TamilNadu, India

1976) and also there are many tools that are commercially available with the support of web data mining and web usage mining.

In the business side dataset plays a vital role to get the correct data at the correct time. It is the success of the best business intelligence (Krishnapuram *et al*., 2001). So the arrival of the web data mining is the most welcomed tool for the business intelligence people at the right time. Using this tool most of the companies and the enterprises had get lot of benefits to achieve their goals and tasks and it also overcome the advanced technologies in the data mining (Makkar *et al*., 2012).

After the introduction of the World Wide Web, the user visiting the web page has increased. Whenever the user visits or searches the web, the information was stored in log files. To analyze that data, we introduced data mining process, to extract the data and to find the frequent pages visited by the user (Shaikh *et al*., 2011). The main achievement of web usage mining is attracting the users, analyze the data and implement the data. The data mining in web log is a recent technology which is used in data mining technology to find the browsing patterns and analyze the data (Koh and Yo, 2005).

Further we discuss related work proposed in this area, proposed methodology, results, discussions, conclusion in this study.

1.1. Related Work

The process of web mining is to implement the knowledge discovery and retrieves the necessary information from the World Wide Web (Latif *et al*., 2010). Our main goal and our work are in the area of Web Usage Mining. We have faced many problems in the process of retrieval of data and these problems cannot be rectified correctly in the previous work. So with our advanced research we rectify the problems in this study and here we proposed a new algorithm based on sequential pattern clustering with the help of the Euclidean Distance algorithm. This algorithm is very essential for web usage mining.

In data mining technique we have one of the best methods in the Web Usage mining that is used to find out the patterns of the web data, which is used to serve the best response and results and to fulfill the Web based usages (Bayir *et al*., 2009). To strengthen the web based applications which we can use in all areas like e-commerce and in the web based usage area to give the good solution to the user and also it can analyze all the logical and tough tools to be developed in web usage mining.

In this concept we used path extraction using sequential pattern clustering algorithm. In our web usage mining area many researchers have proved that their databases must be efficient from the Ref of "Similarity measure to identify only user's profiles in web usage mining" (Latif *et al*., 2010). By this concept they identify the user profile details using web usage mining. Data mining also used in e-commerce mainly proved from the concept with the Ref of "A Comprehensive Survey on Frequent Pattern Mining from Web Logs" (Chauhan and Jain, 2011). They rectified the best result by web logs files as an efficient data from the clustering of web log data and the data are analyzed according to the user by the Ref of the" Clustering of Web Log Data to Analyze User Navigation Patterns" (Shrivastava *et al*., 1976). Krishnapuram *et al*. (2001) using relational and clustering methods they prove that the "Low-Complexity Fuzzy Relational Clustering Algorithms for Web Mining" they stated that the data can cluster the log files and they find the accurate result. Koh and Yo (2005) they described the method to analyze and mine the patterns using association rules from "An Efficient Approach for Mining Fault-Tolerant Frequent Patterns based on Data mining with Association rule Techniques".

Anitha and Krishnan (2011), they described a e-learning recommendation framework using rough sets and association rule mining. Bates *et al*. (2010), a Regular expressions considered harmful in client-side XSS filters.

Grauman *et al*. (2008) proposed a hub number of a graph for the inference to identify the most visited hub.

Koh and Yo (2005), fault-tolerant appearing vectors are designed to represent the dstribution that the candidate patterns contained in data sets with fault-tolerance. They proposed VB-FT-Mine algorithm which applies depth-first pattern growing method to generate candidate patterns. They use vector operations on bit vectors to identify a fault tolerant frequent pattern.

Rathamani and Sivaprakasam (2012), a web usage mining and behaviour analysis is discussed. They used fuzzy c means clustering is used for analysis and used real world web log data. Bayir *et al*. (2009), a Framework for Mining Large Scale Web Usage Data is proposed, which uses Map/Reduce paradigm.

From the below **Fig. 1** we had introduced the general web usage mining outline. Mostly we have two architectures in web usage mining structure, first it contains domain dependent process where we can transform the data and second it contains domain independent process, where the transaction process can done. In the first process, we can use normal data preprocessing, data discovery and the outline of the data discovery.

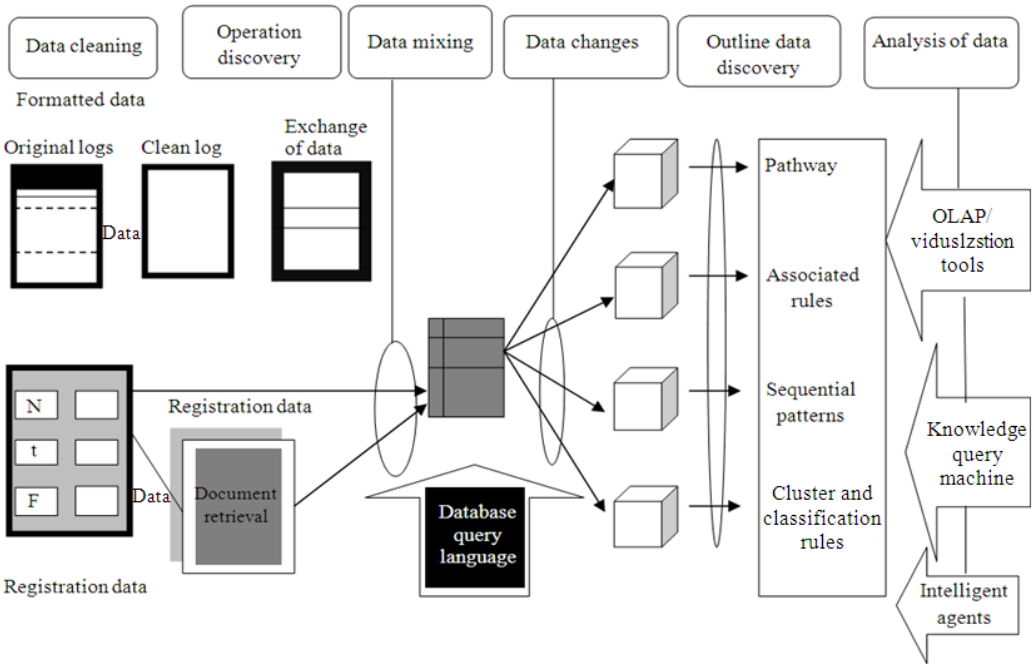

Fig. 1. The web usage mining outline

Here we used the original log files which are taken for the sample data set as an input. The browsing pages which are used by the user may be single page or multiple. After the data cleaning, the log files are clustered and extract the data by associated rules **Fig. 2**. Normally, all the above process is done in web usage mining process. Finally, a query mechanism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints.

For powerful data mining, query languages need the process of data mining tools and systems. The data mining tool which is used in our project is user friendly. Here we proposed some ways and guidelines for web usage process, to extract the data from the large data sets which contains web log files. We can highlight our best result and find the exact data. So with the help of our query mechanism we can provide, extract and retrieve the best web result.

1.2. Proposed Work

In the web usage mining we used to extract the data according to the knowledge by the web pages which were frequently used by the user and here we used path extraction method to extract the content path and transaction path of the user. We introduce the algorithm of Path extraction using sequential pattern clustering methods. The methods which we used are Data preprocessing and data cleaning methods. The discussion is as follows.

2. MATERIALS AND METHODS

2.1. Data Preprocessing

This data pre-processing is very difficult task and it is the major usage in web mining process. This technique is used to select the most needed data's and remove the unwanted data in the log files. This method is used to format the log data and extract the original - information of web user log files.

2.2. Information Cleaning

The study of data cleaning processes to eliminate the mismatches of data or immaterial data. Analyzing the enormous amounts of account in server logs is an awkward activity. We showed our cleaning process in our experiment:

- For the condition based removing, status code has been used. If each record has the status as 300 means the blow level of status code were removed
- The URL file value is checked for the file extension. If the file format require image file or CSS and so on they are removed

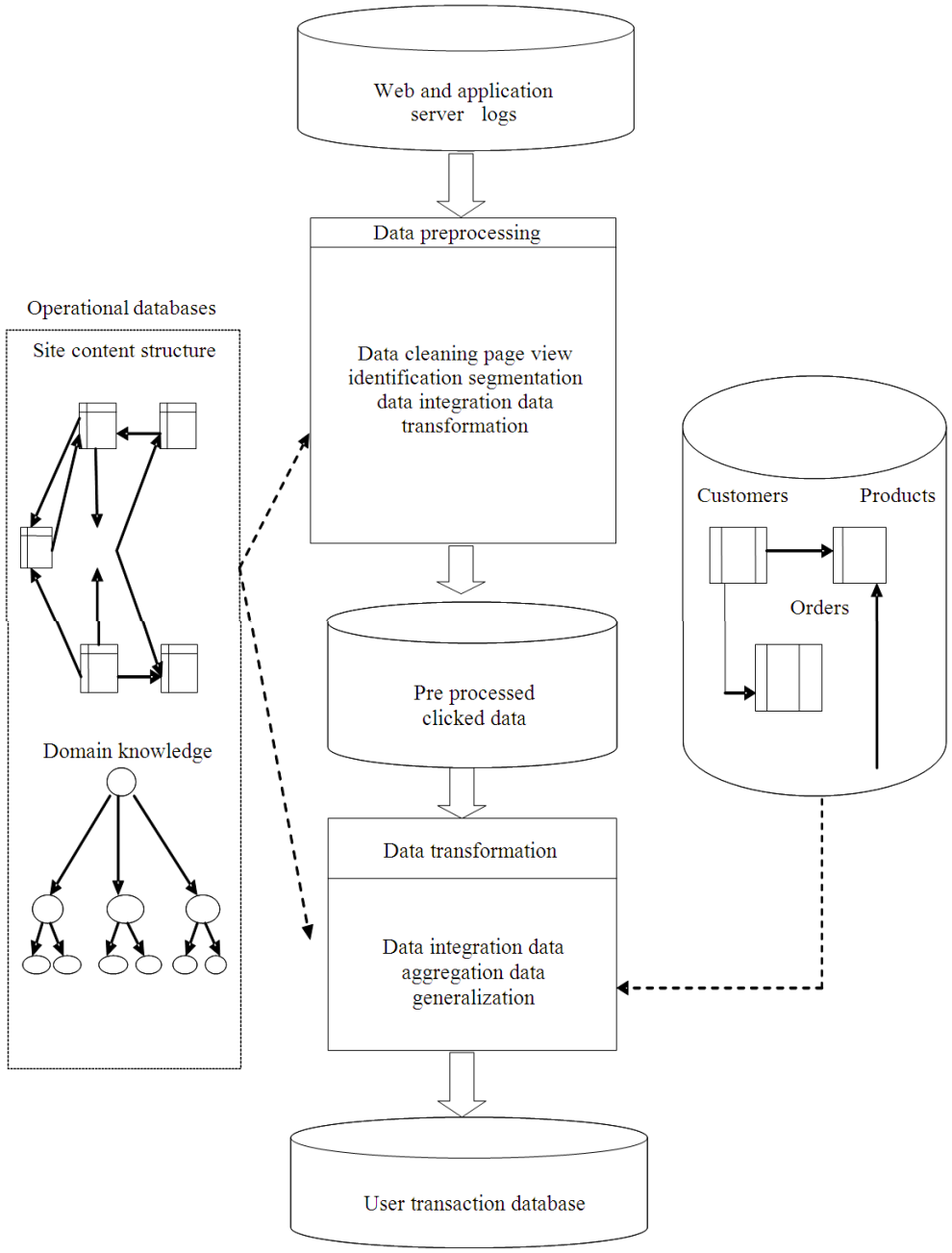

Fig. 2. Proposed web usage mining architecture

Our proposed method is a combination of content path set and travel path set identification. The first step is splitting the forward reference paths by using the maximal forward reference paths in a particular session by user and travel path is also identified from those paths.

Steps to be followed in executing of Analyzing web log files by Euclidean Distance Algorithm:

Start:

1. Input: Log Files. (LF)
2. Output: Elaborated Log File (ELF)
3. Study the Records in the Log files.

4. Understand the each record in the log file.
5. Check the condition if the Status = ”800” and the technique = ”GET” then
6. Show the Client IP address and URL stem.
7. If the contents in the URL = {.zip, .docx, .jpeg, .gif, .jpeg}
8. Remove the suffixes of the URL.
9. Store the IP address and URL.

End if
Continue with checking of next record
End

2.3. User Identification

The log files data after the cleaning is considered as Web Usage Log Set. The below mentioned algorithm for proposed work is Path Extraction for Web log file by Euclidean Distance Algorithm:

1. Start
2. Construct the calculation of the records and calculate the length of the data, path identification and the user session discovery are extracted using Euclidean distance method and clustering model.
3. The path is extracted by its association rules as:
4. Identifying the path of the User: User ID, URI1, Date1 and RefLength1)…(URIk, Date K, RL length)
5. Data Cleaning: {User IP, Date, Method, URI, Status, Version, Bytes, Referrer URL and Browsers OS}
6. Length of the data is calculated by: The length of the pages used by the user with its times= number of the bytes are used by the user.

Where:
Ref LT’ = Time difference between a record and next record.
Send data (bytes) = log data taken from entry
c = Transfer rate

7. Process of User Session detection:

If URL in current record
Not accessed previously
Or
IF referrer URL field == empty
{
Consider as New session
}

2.4. Referrer URL

Users and sessions are identified by using these fields as follows. If two records has same IP address.

Reference Length defines as the time taken by user to inspect a particular page. This length is act as main role for mining and clustering of user activity.

In our process, we calculate the length according to the information which is gathered by the user browsed pages and this process, will give us exact time, because the loading of web pages can take enormous time to get the transfer rate from the internet and it can load all types of files, but only the multimedia files can take more time to load. We can calculate the length according to the loading time and the length of the page.

2.5. User Session Discovery

The task of session discovery is to divide the accessed pages according to the user browsed in the Web. Those divided sessions are taken as the parameter value to classify, forecast and cluster the data into groups:

User’s session = {UserID, (UResourceI1, ReferrerURI1 and Date1)….. (URIm, Referrer URIm, Datem))}

Here $1 \leq m \leq n$, n total of account in a user logs. Each record in user log has to belong to a session and each record in user logs can be in the right place to one user session only. After alignment the account into sessions the path completion step follows.

3. RESULTS

For our results in this implementation we take the sample log of our website which is to be retrieved to extract the data.

In **Fig. 3**, According to the length of all records which are used in the particular web site we can identify the users and sessions. So mainly we used the two methods in the proposed system which is Travel path transactions and Content path set.

In **Fig. 4**, we used data set for preprocessing to extract the patterns according to the user viewed in the same website. The preprocess method is done by the process of classification and association rule data mining to extract the path, which the viewing can be done easily. After preprocess method we can get the IP address and the method which should be GET or POST to extract the record from the log file.

In the **Fig. 5**, we gave the clear details of the user profile according to the IP address of the user and the number of bytes with the time and date of the user. This can be frequentely viewed by the user.

From the **Fig. 6**, we can analyze the data by graph method according to our experimental result. The process was used to find out the number of pages used by the user with the completed path and the session.

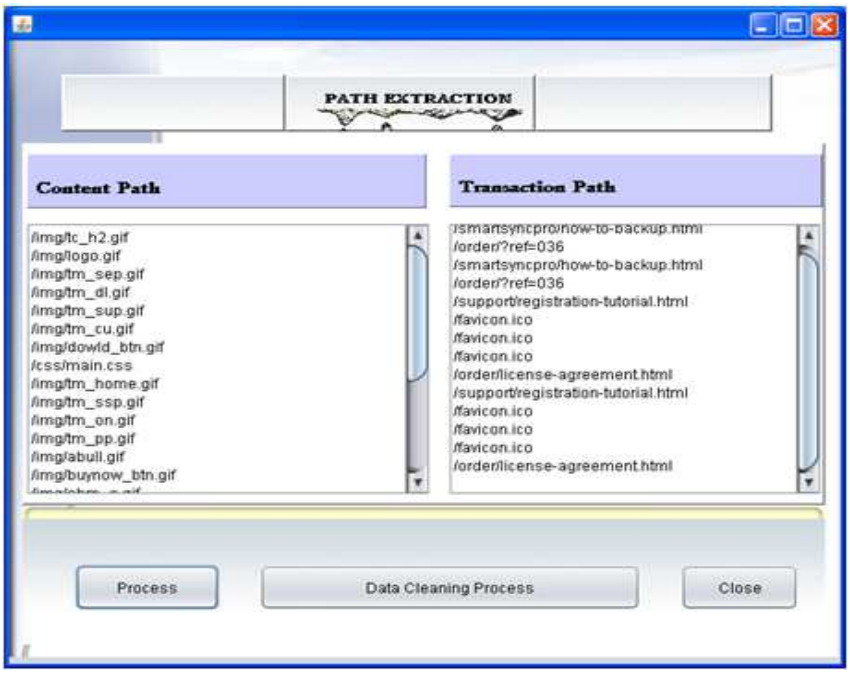

Fig. 3. Path extraction on log files

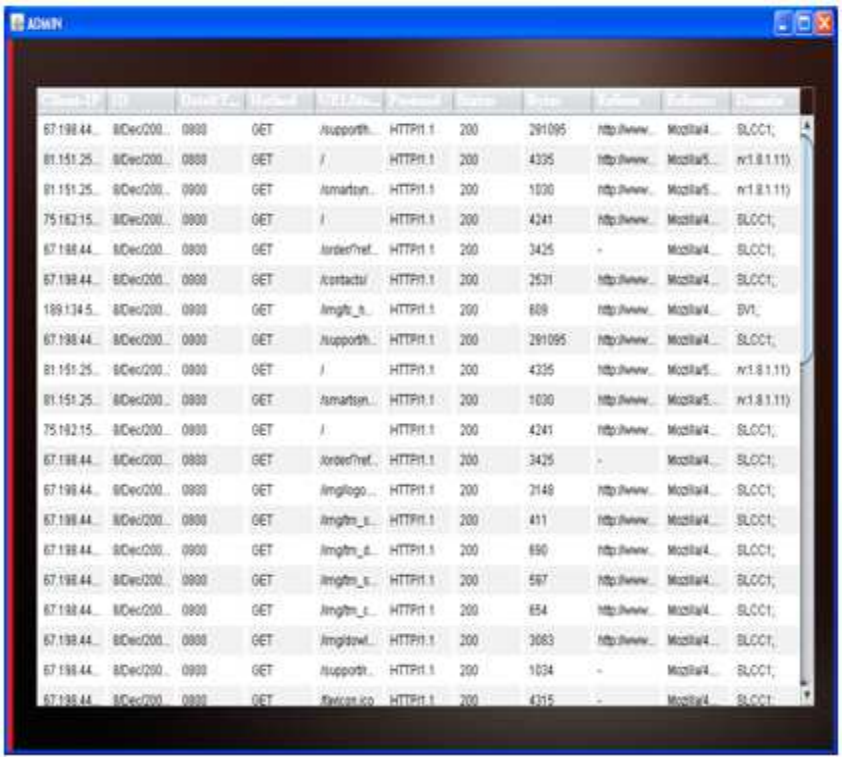

Fig. 4. Dataset of after preprocessing

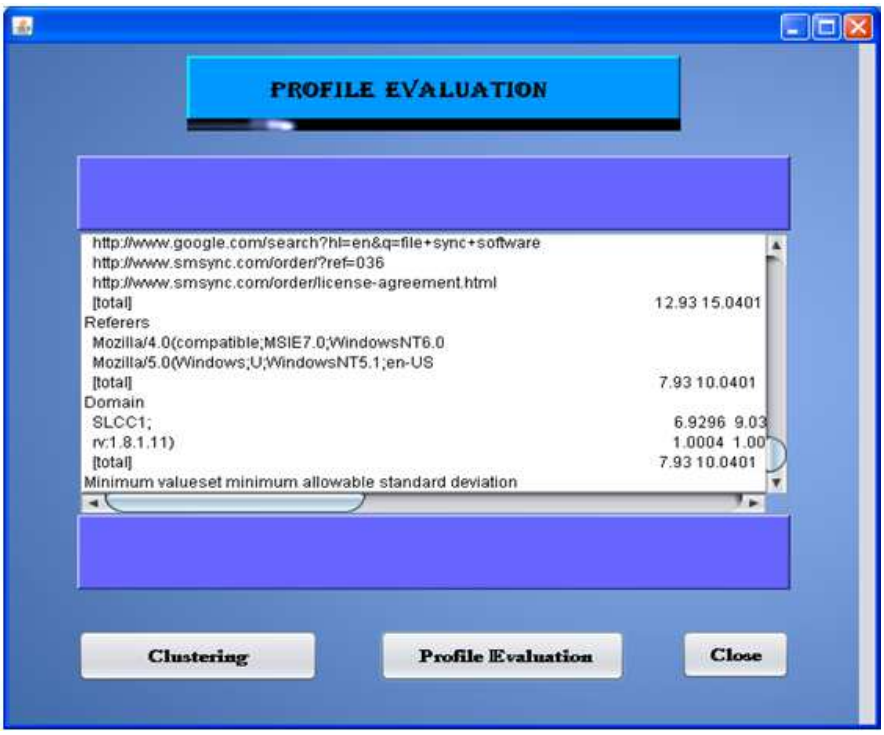

Fig. 5. Evaluation of user after clustering

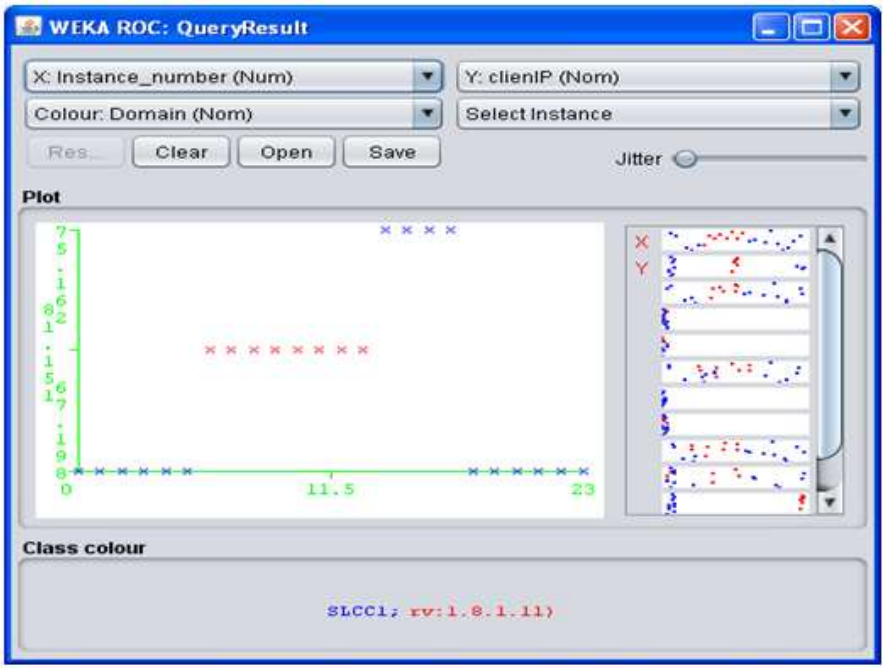

Fig. 6. Result analysis of the log files using graph

Table 1. The path completed by user after pre-processing

User IP address	User ID	Session ID	Path completed
192.168.100.24	2	3	1-10-14-18-10

Table 2. Most interest path by user

User IP address	User ID	Session ID	Path completed
192.168.100.24	2	3	1-3-10-14-17-18-3-10
192.164.135.102	4	2	3-5-6-10--18-21-16

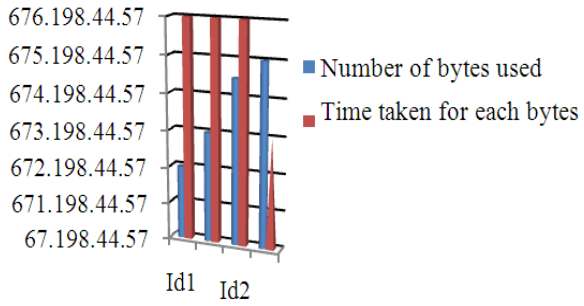

Fig. 7. Number of bytes between users

The **Table 1** shows the completed path of a user for a log file using path pattern discovery using a user website. So the calculating process is done using the knowledge discovery pattern mining to extract the path from the original web server log files. The path completed of the client ID is extracted using the path and the length of the records according to the path calculation.

The below mentioned table will give the detail process about the IP address and the path completed by the user.

The **Table 2** shows nearly all the relevant pages which are used by the user to calculate the original data. After the data preprocessing method the user ID and the Session ID of the attributes, the data cleaning process the log files were applied to calculate the reference length.

4. DISCUSSION

Figure 7, shows the frequency of users from different IP's according to the number of bytes.

The proposed algorithm makes more difference than other algorithms in identifying the web page frequency or to extrack knowledge from web log files. Eventhough the web log files are huge in size, the algorithm we proposed will be efficient in extracting knowledge from the log files. The extracted knowledge can be used for anything like identifying the web page frequency or to infer some useful knowledge from web user sessions.

5. CONCLUSION

With the explosive development of the web mining and its usage for the World Wide Web we have an opportunity to analyze Web data and extract all the data and analyze the discovery knowledge from it. The past five years have seen the emergence of Web data mining as a rapidly growing area, due to the efforts of the research community as well as various organizations which are practicing it. The study which we have proposed here will give detail description of web log file, with its IP address, Name of the user, Time and date and the contents of the users. Added this information it proposes detail conditions that to be followed in the web usage mining process. Here with the different mechanisms we perform and extract large log files.

6. REFERENCES

Anitha, A. and N. Krishnan, 2011. A dynamic web mining framework for e-learning recommendationsusing rough sets and association rule mining. Int. J. Comput. Appli., 12: 36-41. DOI: 10.5120/1724-2326

Bates, D., A. Barthm and C. Jackson, 2010. Regular expressions considered harmful in client-side XSS filters. Proceedings of the 19th International Conference on World Wide Web, Apr. 26-30, ACM Press, New York, USA., pp: 90-100. DOI: 10.1145/1772690.1772701

Bayir, M.A., I.H. Toroslu, A. Cosar and G. Fidan, 2009. Smart Miner: A new framework for mining large scale web usage data. Proceedings of the 18th International Conference on World Wide Web, Apr. 20-24, ACM Pres, New York, USA., pp: 161-170. DOI: 10.1145/1526709.1526732

Chauhan, A.S. and S.K. Jain, 2011. A comprehensive survey on frequent pattern mining from web logs. Int. J. Adv. Eng. Appli., 1: 138-142.

Grauman, T., G.S. Hartke, A. Jobson, B. Kindersley and D.B. West *et al*., 2008. The hub number of a graph. Inform. Proc. Lett., 108: 226-228. DOI: 10.1016/j.ipl.2008.05.022

Koh, J.L. and P.W. Yo, 2005. An efficient approach for mining fault-tolerant frequent patterns based on bit vector representations. Database Syst. Adv. Appli., 3453: 568-575. DOI: 10.1007/11408079_51

Krishnapuram, R., A. Joshi, O. Nasraoui and L. Yi, 2001. Low-complexity fuzzy relational clustering algorithms for web mining. IEEE Trans. Fuzzy Syst., 9: 595-607. DOI: 10.1109/91.940971

Latif, F.A., D. Nicolas, M. Nicolas and P. Jean-Pierre, 2010. Similarity measure to identify users' profiles in web usage mining. Universite de Rouen.

Makkar, P., Seema and K. Aggarwal, 2012. Comparison of Pre-fetched Pages Before and After Path Completion. Int. J. Comput. Sci. Eng., 4: 1420-1426. PMID: 103505441

Rathamani, M. and P. Sivaprakasam, 2012. Cloud mining: Web usage mining and user behavior analysis using fuzzy C-means clustering. IOSRJCE, 7: 9-15.

Satish, B. and S. Patil, 2011. Study and Evaluation of user's behavior in e-commerce using data mining. Res. J. Recent Sci., 1: 375-387.

Shaikh, Z.F., G. Kulsundar, M.S. Shrivastava, V.V. Ramteke and S. Yadav *et al*., 2011. Pancreatic ascites in the setting of portal hypertension. BMJ Case Reports. DOI: 10.1136/bcr.08.2010.3271

Shrivastava, S., S.M. Tadavarthy, T. Fukuda and J.E. Edwards, 1976. Anatomic causes of pulmonary stenosis in complete transposition. Circulation, 54: 154-159. DOI: 10.1161/01.CIR.54.1.154