International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 8, Volume 2 (August 2015) www.ijirae.com

Pre-Processing and Analysis of Web Server Logs

Chaitra L Mugali
Department of Computer Science, KLS Gogte Institute of Technology Belagavi, Karnataka, India

AyeshaAzeema Maniyar
Department of Computer Science, KLS Gogte Institute of Technology Belagavi, Karnataka, India

Asst. Prof. Padma Dandannavar
Department of Computer Science, KLS Gogte Institute of Technology Belagavi, Karnataka, India

Affiliated to Visveswaraya Technological University, Belagavi

Abstract— Nowadays, World Wide Web has become a huge repository or storage for retrieving, storing, sharing and also distribute the data or information. Web is a dominant platform from where the knowledge can be discovered to study web user behaviour. Each and every interaction of user with the web will be recorded or stored in a text file which is basically called as Web Log File. These web log files will be in ".txt" format. The data which is stored in web log files will be consisting of huge amount of information with some kind of incomplete and unwanted data too. It is difficult to deal with whole data which is huge in size. So, unwanted or uninterested data can be removed by processing the data. The process of applying the data mining techniques on web data to discover the interesting patterns is known as web mining. Three kinds of web mining are namely web usage mining, web structure mining and web content mining. Here in this work web mining is done with web log files for studying user interactions with web or user accesses. This is known as web usage mining. The three important steps or modules in web usage mining [1] are, Web logs pre-processing, Patterns discovery, Patterns analysis. The result of web usage mining can be used for various purposes like "improving Web Design", "Web Personalization", "E-commerce", "Pre-fetching Web Contents", etc., thus helping customer satisfaction.

Keywords— Pre-processing, Pattern Discovery, Pattern Analysis, World Wide Web (WWW), Web Log File, Web Usage Mining (WUM), Web Log Mining

I. INTRODUCTION

World Wide Web has become a huge repository or storage for retrieving, placing, sharing and also distribute the data or information. Web is a dominant platform from where the knowledge can be discovered to study user behaviour. In the era of internet, usage of web applications and the number of web users are getting increased enormously very high. Web has become most popular platform for attracting and satisfying the users.

Each and every interaction of web user with the web will be recorded or stored in a text file which is basically called as Web Log File. These web log files will be in ".txt" format. Each interaction of user with web or server will be recorded as a single record in web file containing logs. These server log files are used to understand or study behaviour of the web users. The data which is stored in web log files will be consisting of huge amount of information with some kind of incomplete and unwanted data too. Kindly it's little difficult to deal with whole data which are in huge amount of size. So, unwanted or uninterested data can be removed by processing the data. Data mining techniques can be applied on the web log files to remove out unnecessary data and then finding patterns out of pre-processed data for analysing the data to study web user behaviour. The process of applying the data mining techniques on web data to discover the interesting patterns is known as web mining. Three kinds of web mining are namely web usage mining, web structure mining and web content mining. Here in this work web mining is done with web log files for studying user interactions with web or user accesses. This is known as web usage mining. The three important steps or modules in web usage mining [1] are,

1) Web logs pre-processing 2) Patterns discovery 3) Patterns analysis.

A. WEB LOG FILES

Web log files are the text files which get generated whenever there is a interaction between user and the web. Each user interaction with web will be recorded as a single record in the web log file. Generally web log file records contains fields such as IP address, URL accessed, time stamp, number of bytes, method used for making request and protocol details. These web log files can be used to understand or study the web user behaviour. The data which is stored in web log files will be consisting of huge amount information with some kind of incomplete and unwanted data too. Data mining techniques can be applied to on web log files to remove out unnecessary data and then finding patterns out of pre-processed data for analysing the data to study web user behaviour.

A sample web log record is shown below,

123.46.7.79.8 - [12/Mar/2012:04:06:50 -0500] ―GET/HTTP/1.0‖ 200 3240

Where,

- 123.46.7.79.8- IP address
- "-"(hyphen) indicates Anonymous user id
- 12/Mar/2012:04:06:50- Web page access time
- -0500- The time zone
- GET/HTTP- HTTP request method
- 200- HTTP status code
- 3240- Number of bytes transmitted

1) WEB LOG DATA SET:

The dataset used in this work are the web log files. These are the web log files which are generated in accordance with interaction of web user with server or web. Each record line in web log file indicates an interaction between user and server. There are different kinds of web log files, which store these kinds of automatically generated log data. Those are referrer logs, access logs, client-side cookies and error logs [3]. Here for this work the web server logs are collected from the website ftp://ircache.net, where the sample web log data sets are freely available to process on them. These Web logs can be downloaded and used for experimenting.

2) TYPICAL SOURCES OF DATA:

Web log files that are automatically generated when there is an interaction between user and server. Web data is available from different sources from web. Few sources are listed below,

- E-commerce websites data.
- Web user profiles.
- Content and structure from the web page.

3) KINDS OF WEB LOG FILES:

Here are different types of web log file stores [2], [5] where web log data will get stored.

a) Web Server Logs (Server Side)
b) Proxy Server Logs (Proxy Side)
c) Browser Logs (Client Side)

a) Web Server Logs (Server Side):

Web server logs are stored on server side. Web server logs will be containing information like IP address, requested URL, time stamp, number of bytes, method and protocol used and etc.

b) Proxy Server Logs (Proxy Side):

Proxy server logs will be stored on proxy server. Whenever main server is unable to response to the user requests, these requests are handled by the proxy servers so at this time proxy server logs will get generated. These proxy server logs will be containing some additional information related to proxy server in addition to as that of web server log files.

c) Browser Logs (Client Side):

Browser log files are stored on client machine. These browser log files will be containing more clients related or user specific information in addition to web server log files information.

4) DIFFERENT WEB LOG FILE FORMATS:

The web log files which get generated automatically from user interaction with server are stored in different formats. Those are,

a) Common log file format
b) "Microsoft IIS (Internet Information Services)" web log format

a) Common Log File Format:

Common web log file format is standard text log file format used to store web log records. Each request or the user interaction will be stored as a single line record with comma separated fields in web log file.

b) Microsoft IIS Log Format:

This log file format is non-customizable format of ASCII which is used for storing web log records. It stores more fields or information compared to NCSA file format but less than Common log file format. This is also a standard format of storing web log records.

B. WEB MINING

Web mining process can be defined as the process of applying data mining algorithms or techniques on web data so as to discover the interesting or access patterns or knowledge to study user behaviour or user access patterns. Web mining process is categorized as "web usage mining (i.e., WUM), web content mining and web structure mining.

a) Web Content Mining:

Web content mining deals with mining web pages content like text, images for finding out the content relevance to the search query. It helps to analyse which content is important and which is not and also what is the proper position to keep important content.

b) Web Structure Mining:

Web structure mining of web data deals with mining or identifying relationships among different web pages which are linked by direct links or by information. It helps to generate summaries of web pages in structured way.

c) Web Usage Mining (WUM):

Web usage mining deals with mining web user accesses. This helps to discover user access patterns of accessing web pages. User accesses will be stored in Access Log Files. These web access log files are used for web usage mining.

C. WEB USAGE MINING (WUM)

Web usage mining is the process of mining web data, in this case the web data considered is web log data or files. Here web usage mining is performed to analyse web server log files. This is done to discover frequently accessed web pages flow pattern. This is popular research area in mining of web. This is mainly focuses on studying behaviour of web users and also web user's interactions with web server. This includes finding or discovering frequent item sets or association rules out of given web log data.

Process of Web Usage Mining:

The three main steps of web usage mining namely web data pre-processing, web pattern discovery and web pattern analysis.

1) Web Data Collection And Pre-Processing:

Data collection is the foremost step in web usage mining where in the web log data is collected. Web log data is taken from an available website ftp://ircache.net. Each record line in this web log file represents a user interaction with the web server. Each record line in this web log file contains the fields like time, duration, client address, result codes, bytes, request method, URL, rfc931, hierarchy code, type.

Sample record:

1168300919.015	1781	17.219.121.198	TCP_MISS/200	1333	GET
http://www.quiethits.com/hitsurfer.php?[387.vlWHJccIgu:aIqlvda] - DIRECT/204.92.87.134 text/html					

This collected web server log files will be containing huge amount of log records with some kind of unwanted data too. Kindly it's little difficult to deal with such huge amount of web log data. So this kind of unwanted or unnecessary data has to be removed out before proceeding to next step. This is done in pre-processing step by applying different pre-processing techniques. Some of the pre-processing techniques [8], [16] are "Data cleaning", "User identification", "Session identification", "Data transformation", "Path completion".

2) Web Patterns Discovery:

Web patterns discovery step is performed to discover interesting patterns or knowledge to analyse web user behaviour. Once the pre-processing step is completed the patterns or knowledge can be discovered from the pre-processed web log data in pattern discovery step. Different methods or techniques are used to discover association rules or frequent patterns like "statistical methods" and also data mining methods like "Path analysis", "Association rule", "Sequential patterns", "Clustering" and "classification" [22]. These are performed on web log files so as to detect interesting patterns to study web user behaviour. These discovered patterns or knowledge can be represented in some form like table, graph and charts etc. Algorithms like FP-Growth and Apriori are used in this phase.

3) Web patterns analysis:

Web pattern analysis is the process in which uninterested patterns are removed out from patterns discovered in previous pattern discovery step. Here the patterns discovered are analysed by making use of some of OLAP tools or by SQL query mechanism.

II. RELATED WORK

In today's world as the usage of internet has increased enormously, there is the need for understanding each and every web user behaviour so as to improve the business and to satisfy user expectations. So the research on Web Usage Mining by many of the researchers and expertise and analysts is going on.

In [5] Vijayashri Losarwar and Dr. Madhuri Joshi have proposed a methodology of pre-processing steps like Data cleaning, User Identification, and Session Identification mainly for Web Personalization.

Data cleaning includes removing unwanted fields of web log records, eliminating the records with file names like .gif, jpeg, jpg, java script, css, robot.txt, etc, and also removing the records with failed HTTP status code. User identification involves identification of user by assuming each combination of IP address, Agent and Operating System as a single user. In session identification the session is defined as the combination of pages accessed with time given. Like 'S' denotes a session. Session 'S' indicates the set of pages accessed by a particular user.

Sheetal A. Raiyani and Shailendra Jain [6] have explained the web log pre-processing techniques which include Data Cleaning, User Identification, Session Identification and Path Completion. Data cleaning includes removing out the data which is not required like the records with failed HTTP status code and with jpeg and jpg, css file extensions. User identification and session identification is done with some of the rules given for identification. Path completion is done to fill up the missing page references by checking the referrer logs.

Two approaches for Web log pre-processing are explained by Ms. Dipa Dixit and Ms. M Kiruthika in [7]. One approach is achieved using XML and other is achieved using text file, but basic steps involved in pre-processing are same for both. Web log pre-processing approach using XML gives structure to the web log records of web log file. This provided XML structure makes it easier to understand the web log file. And another pre-processing approach using text file is done by separating each field or attribute of the record by a delimiter as a space, this helps to differentiate each attribute from other.

Surbhi Anand and Rinkle Rani Aggarwal [9] have explained the Data field extraction and Data storage for web log files. The field or attribute separation is done by making use of delimiter as a space. They have implemented a data field extraction algorithm using Java language by using java inbuilt methods. The data storage for storing the logs after the data field extraction is done by creating log table by making use of SQL query.

As is known web personalization is common application of data mining, the techniques used to achieve this are content based filtering, collaborative filtering, rule based filtering. In [10] the authors have explained personalized collaborative filtering method by combining with association rule mining. They have built a framework for Web personalization. And they have also discussed the drawbacks of both Apriori and FP growth algorithms.

Nanhay Singh, Arvind Panwar1, and Ram Shringar Raw [4] have proposed a framework to improve the web proxy server performance by making use of web usage mining and pre-fetching technique. Authors have proposed the framework for predicting the web user requests and then pre-fetching the content from the server. Authors have compared the analysis results using LRU and LFU algorithms considering hit ratio and byte ratio. They have carried out simulations by implementing the framework in C using MATLAB.

Ramya C, Dr. Shreedhara K S and Kavitha G [13] have proposed a pre-processing methodology for processing web logs in order to discover the interesting patterns out of it. Here the complete pre-processing methodology involves Merging, Data Cleaning, User/Session Identification, Data formatting and Summarization. This proposed pre-processing methodology has proved that it reduces the size of initial web log data efficiently and increases the quality of data.

The Data Pre-processing and Data transformation techniques are explained by Abdul Rahaman Wahab Sait, and Dr.T. Meyappan in [14]. Authors have mainly concentrated on pre-processing of web logs and then converting the data into numerical form which will become very easy to generate interesting patterns. As we know web logs stores the IP address of the user who has accessed a page this violates the rule of not disclosing the user details to the outside world so this data transformation helps to hide the user details from outside world.

Web Usage Mining involves Pattern Discovery and Pattern Analysis on the web log data which is generated by the user interaction with the web. Mr. Rahul Mishra and Ms. Abha Choubey[20] have worked on discovering the interesting patterns from the web log records by making use of FP-growth algorithm. Authors have made use of FP-Growth algorithm to discover the interesting patterns or frequent item sets. They have also given the performance comparison between FP-Growth algorithm and Apriori algorithm. FP-Growth algorithms make use of FP-tree data structure. The comparison here showed that FP-Growth is the best algorithm for mining frequent item sets.

Nanhay Singh, Achin Jain, Ram Shringar Raw [21] has done comparison analysis work on web usage analysis by making use of some of the pattern recognition techniques. The pattern discovery techniques used by authors here are Converting IP address to Domain name, Grouping, Filtering. These pattern recognition techniques help to recognize the patterns. The Bandwidth/Hit Usage comparison has been done for image files by authors. This helps to improve the websites performance indirectly by improving structure, content, delivery and presentation.

III. SYSTEM ARCHITECTURE

System architecture is a type of conceptual design of a given system which includes system structure and behavior. This description of architecture gives explanation about system and its properties.

The system architecture for web usage analysis is shown here. There are three main modules in the system namely:

A) Data Acquisition and Pre-processing
B) Pattern Discovery
C) Pattern Analysis

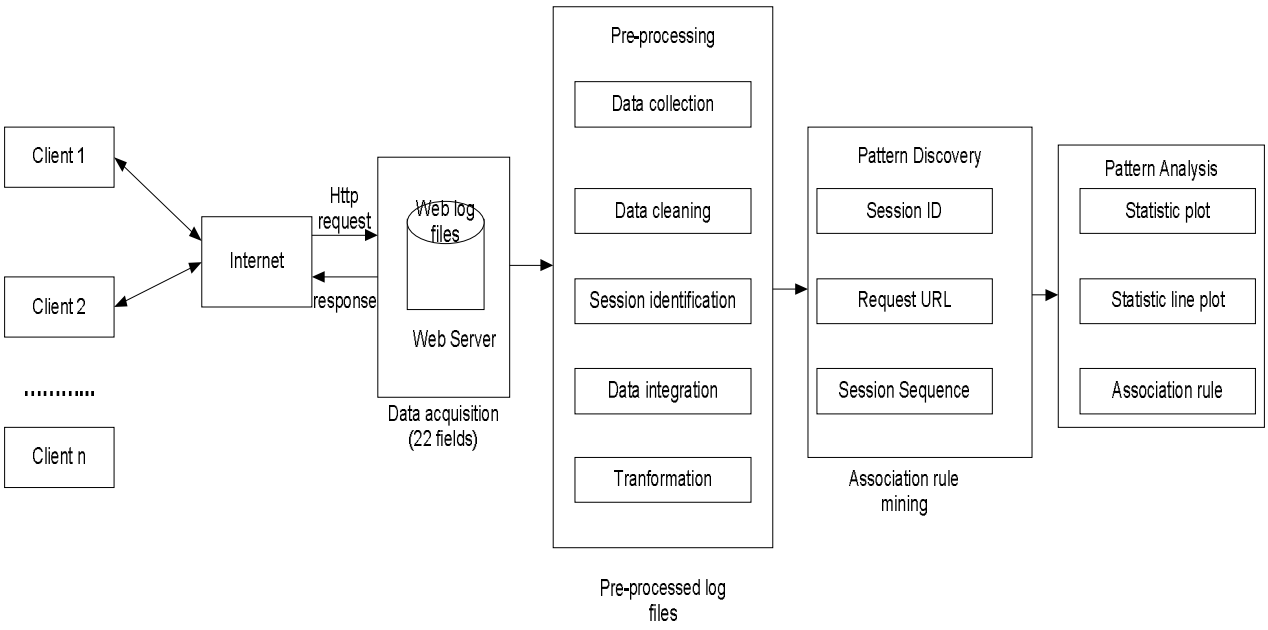

Fig 1. System Architecture

A) Data Acquisition and Pre-processing:

The web data set selected or considered for this work is web log data which is taken from ftp://ircache.net. Applications of web usage are mainly based on web data collected from sources like,

a) Web servers
b) Proxy servers
c) Web clients

The steps or methods involved here in web data pre-processing are [17], [18] [19],

1) Data Cleaning:

Data cleaning is the process of removing out the irrelevant or unnecessary data which is not useful for future work in mining [5]. Here we are interested in files or pages or URLs accessed by web user with file extension ".html". So we are just going to remove out all other web log records which having URLs with file extension other than ".html" like of "gif", "jpeg", "css" and so on . After this step the web log records with file extension ".html" are only used for further process.

2) User Identification:

User identification involves process of identifying the web user. Users can be identified with unique IP address, Browser used also referring operating system used by the web users. Each user is identified individually.

3) Session Identification:

Session identification is the step where in the sessions are identified. Here in this work each of the web log record is taken as a session.

4) Data Transformation:

Data transformation is the step in which the data will be transformed from one form to another which is a relevant form to work on. The data can be transformed to the simplest form which makes easy process to deal with it. In this flow it's kindly little difficult to deal with long URL form, each of the URL is transformed as a unique number. So that it will become easy to deal with each URL with that unique number in future process.

B) Pattern Discovery:

Pattern discovery step is performed to discover interesting patterns or knowledge to analyse web user behaviour. Once the pre-processing step is completed the patterns or knowledge can be discovered from the pre-processed web log data in pattern discovery step. Variety of techniques used to discover the web patterns are like "statistical methods" and also "data mining methods" like "Path analysis", "Association rule", "Sequential patterns", "Clustering" and "Classification". These methods are applied on web log data so as to study web user behaviour. These discovered patterns or knowledge can be represented in some form like table, graph and charts etc. The data mining algorithms can be used here to discover the frequent item sets or patterns. Here in this work two data mining algorithms namely FP-Growth and Apriori are used to discover the frequent item sets.

1) Statistical analysis:

Statistical analysis methods can be used to discover the frequent patterns or knowledge about the web users. The different types of statistical analysis like mean, median frequency. These can be used on web logs to discover the page access knowledge out of it.

2) Association Rules:

The generation of association rules out of given web log data using web mining techniques is done here. In the web usage mining domain the association rules will refer to the set of web pages which are accessed together and also the set of web page patterns which are accessed very frequently by the web users. It also refers to the pages which are referenced together.

3) Clustering:

Clustering is the web mining technique which is used to group together the set of items into different clusters having similar characteristics or features. In web usage mining the types of clusters can be created are say page clusters which includes set of pages grouped with respect to web page content and next is usage cluster which includes the pages which are grouped together with respect to web page usage.

4) Classification:

Classification is the data mining technique which is used for mapping of a given data item into one of many predefined classes. For example, developing the profile of web users belonging to any particular class. It is a supervised learning. The algorithms like “decision trees”, “naive bayesian classifiers”, “k-nearest neighbour classifiers”, and “Support Vector Machines” can be used here.

C) Pattern Analysis:

The pattern analysis is performed to remove out the uninteresting patterns or rules froe frequent item sets which are found in previous pattern discovery step. One of the common forms of pattern analysis consists of “knowledge query mechanism” like “SQL”. In general the techniques used here are “visualization techniques”, “OLAP techniques”, “Data & Knowledge Querying”, and “Usability Analysis”.

IV. IMPLEMENTATION

A) PRE-PROCESSING:

The data pre-processing step is performed in order to pre-process the given raw web log files by removing out the unwanted and unnecessary data and it also involves many other tasks for data pre-processing. The pre-processing module involves steps like “Data cleaning”, “Session identification”, “Data integration” and “Data transformation”. Each step is explained below in brief.

1) Data Cleaning:

In this module the unwanted log records which are not useful for the further process are removed from the web log file in Data Cleaning step. Such as the records with ".jpg",".jpeg",".gif",".png","robout.txt","slurp","bot","script","css",".avi" ".js", extensions is removed from the input log file. Then the processed file will be used for the further processes like Pattern Discovery and Pattern Analysis. The number of lines filtered out will be shown in GUI after pre-processing step.

2) Session Identification:

In Session Identification the sessions in web log file are identified for the further process. Here each log record in the web log file is a session in this work. Each record represents a user request to the server.

3) Data Integration:

In Data Integration it takes only the data which is required. Here in this work only the User ID, the IP address from which user is browsing and the URL which he is accessing will be taken into account.

4) Data Transformation:

In Data Transformation the data will be transformed into other form. Later each web log record will be transformed into an equivalent number so as to make work easy as it is difficult to work with full string of request record in web log file.

B) PATTERN DISCOVERY:

Pattern discovery involves discovering the patterns or rules out of pre-processed web log records or web log data. This discovers the series of URL’s accessed by the user. That is sequence of URL’s browsed by each user is discovered.

1) ALGORITHMS USED:

There are ample of data mining algorithms available which can be used for data mining. Some of the popular data mining techniques or algorithms used in this work are K-means clustering algorithm for clustering, Apriori and FP-Growth for pattern discovery. Each of these algorithms is explained below.

a) K-means Clustering Algorithm:

K-means is one of the common data mining clustering algorithms which can be used for clustering or grouping the items having similar characteristics. This solves the problem of clustering. This algorithm is very simple and easy to understand and also it follows a very easy way for classifying the items given. Here the K-means algorithm will group the given data set items into certain number of clusters say “K” clusters each having an associated centroid with it.

Then next step is to take each and every item or point from the data set and then associate it to the nearest cluster. Once after all the points are covered the "K" centroids need to be re-calculated. Then again each and every point is processed and associated to nearest centroid point. This will be continued until no more changes can be done.

b*) FP Growth Algorithm:*

FP-Growth algorithm one of the popular and efficient data mining algorithms which can used for finding the association rules or frequent item sets or interesting patterns. Here in this work FP-Growth algorithm is used to find frequent item sets from given web server log files. This algorithm works in divide and conquers way. FP-Growth algorithm needs to perform only two database scans. In its first database scan algorithm will find frequent item sets which are sorted by frequency and also in descending order. Next in its second database scan the database will be compressed as FP-Tree. Than the algorithm will start finding or generating FP-Tree for each of the item whose support value will be greater than or same as given threshold value.

Here in this work the FP-Growth algorithm is used in web usage mining. "Web usage mining" includes mining web of data that is web server log data in order to find the frequent web access patterns or frequently accessed web pages. This will help to analyse the web user behaviour so as to help improve the performance of web applications and also help improve business. Here the frequently accessed web page patterns are discovered which will help to analyse web usage. By this result we can pre-fetch and keep frequently accessed web page sequence in cache so as to reduce the page access time in future.

c) *Apriori Algorithm:*

Apriori algorithm is a popular data mining algorithm or technique which can be used to finding or discovering association rules or frequent item sets. This process of finding frequent item sets proceeds in the same way finding individual frequent items from the given database and then it go on extending it to bigger and bigger data item sets as those item sets will appear sufficiently in database very often. The frequent item sets or patterns found from this algorithm can be used to analyse the web user trends or interest or behaviour. The popular application of this algorithm is "market basket analysis"

This is a classic algorithm which can be used in data mining domain for purpose of learning association rules.

- What is the use of association rule learning?
- Nowadays, shopping canters or malls make use of association rules to place the items next to each other or together so that users can buy more items.
- In Amazon, they make use of association rule mining to recommend us the items based on the current item we are browsing or buying.
- Next is an application is the auto-complete feature in Google search engine.

Here in this work the apriori algorithm is used for finding the frequent patterns or item sets or frequently accessed web page patterns so as to help pre-fetch the web pages in cache which are accessed together. But unfortunately apriori algorithm is not so efficient as that of FP-Growth algorithm in discovering association rules.

2) COMPARISON OF ALGORITHMS:

The data mining algorithms like FP-Growth and Apriori algorithms are used here in this work. The performance comparison of both the algorithms in web mining or in generating association rules or finding interesting patterns is done here. In general FP-Growth algorithm is known as an efficient data mining algorithm. Here we have considered two parameters to compare the performance of FP-Growth and Apriori algorithms. One of the two parameters is "time taken" for finding interesting patterns or association rules and other one is the "number of patterns" found. FP-Growth algorithm will find large number of interesting patterns but it takes more time than the apriori. And the apriori algorithm will take less time than FP-Growth but it will result in very few numbers of interesting patterns. But FP-Growth is the most efficient data mining algorithm in respect of number of interesting patterns than apriori as it finds more number of interesting patterns.

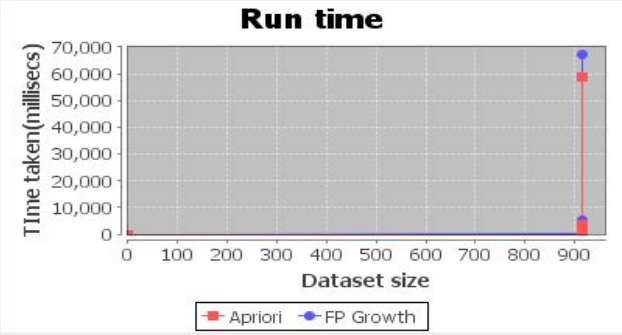

Fig 2. Graph showing time taken by both algorithms

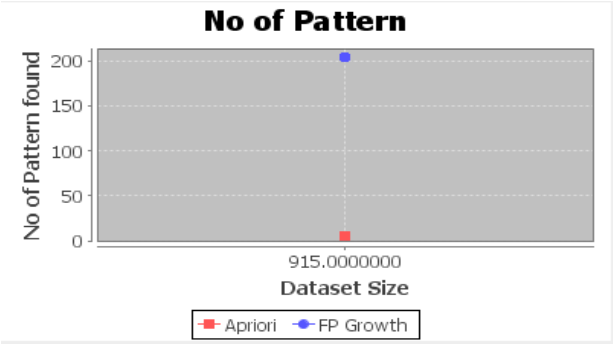

Fig 3. Graph showing the number of patterns found by both algorithms

Here we can see the graphs drawn for showing the time taken by both the algorithms for finding the patterns and also the graph for showing the number of patterns found by both the algorithms. The FP-Growth algorithm is more efficient than Apriori algorithm to discover the interesting patterns as it helps to discover huge number of interesting patterns as compared to Apriori algorithm.

C) PATTERN ANALYSIS:

The pattern analysis is performed to remove out the uninteresting patterns or rules froe frequent item sets which are found in previous pattern discovery step. One of the common forms of pattern analysis consists of “knowledge query mechanism” like “SQL”. In general the techniques used here are “visualization techniques”, “OLAP techniques”, “Data & Knowledge Querying”, and “Usability Analysis”.

The pattern analysis is done for analysing the user access patterns and the user traversing paths of websites. This has many applications in industry. This pattern analysis will help improve the business by studying the user activities and interests. For example finding frequently accessed web pages we can pre-fetch the set of web pages and place in cache which will help user to access those pages in very less time which in turn save the time and satisfy the user requirement.

V. EXPERIMENTAL RESULTS

The web log data or web server log files are processed by applying the data mining techniques to discover the interesting or frequently accessed web pages. Here the sequence of frequently accessed web pages is discovered. This result is used to pre-fetch the sequence of frequently accessed web pages in advance and keep those in cache so as to reduce the web page access time of the pages which are accessed frequently. This way web usage mining help to satisfy the web user. The two popular data mining algorithms named FP-Growth and Apriori are used in this work to discover the interesting patterns. Here we can see that FP-Growth is more efficient than Apriori to discover the interesting patterns as it helps to discover huge number of interesting patterns as compared to Apriori.

Here in following screenshot we can see that the sequence of frequently accessed web pages which can be pre-fetched in cache in advance.

Fig 4. Screenshot showing the sequence of pages to be pre-fetched and cached

VI. CONCLUSION

Web Usage Mining of Web Log Files results into interesting "Rules", "Patterns", and "Statistics". In general the goal of web usage mining is to discover interesting patterns or navigational patterns or knowledge about web users. Here gathered information about web users page access patterns can be used for improving the web applications by pre-fetching and caching web pages. Pre-fetching and caching helps to reduce the page access time or fetching time or latency.

VII. FUTURE SCOPE

Web usage mining technique is used to mine or extract the knowledge from web logs generated on web server. The applications of web usage mining are like improving website design, improving performance of system, pre-fetching and caching.

The knowledge obtained in web log mining can be used for many purposes like [8], [11], [12] in future like,

A) For personalizing the delivery of content of web.
B) To reach customer satisfaction.
C) For improving website design and in E-commerce.
D) For improving the web applications performance by pre-fetching and caching.
E) In internet marketing intelligence.

REFERENCES

[1] Mehak(ME), Mukesh Kumar (Assistant Professor), Naveen Aggarwal (Assistant Professor), Computer Science & Engineering Department, University Institute of Engineering &Technology, Panjab University, Chandigarh, India, "Web Usage Mining: An Analysis", Journal Of Emerging Technologies In Web Intelligence, Vol. 5, No. 3, August 2013

[2] Naga Lakshmi, Raja sekhar Rao, Sai Satyanarayana Reddy: "An Overview of Preprocessing on Web Log Data for Web Usage Analysis", Published in International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-4, March 2013

[3] R. Lokeshkumar, R. Sindhuja, Dr. P. Senguttuvelan, Assistant Professor – (Sr.G), PG Scholar, 3Associate Professor, Department of Information Technology, Bannari Amman Institute of Technology, Tamilnadu: "A Survey on Preprocessing of Web Log File in Web Usage Mining to Improve the Quality of Data", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume-4, Issue-8, August 2014

[4] Nanhay Singh, Arvind Panwar and Ram Shrinagar Rao, Ambedkar Institute of Advanced Communications Technologies and Research, Delhi, India: "Enhancing the Performance of Web Proxy Server through Cluster Based Pre-fetching Techniques", International Conference on Advances in Computing, Communications and Informatics (ICACCA), 2013

[5] Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore

[6] Sheetal A. Raiyani, Shailendra jain, "Efficient Preprocessing technique using Web log mining", 1Department of CSE(Software System),Technocrats Institute of Technology, Bhopal,India; 2Department of CSE, Technocrats Institute of Technology, Bhopal,India, International Journal of Advancements in Research & Technology, Volume 1, Issue6, November-2012 1 ISSN 2278-7763

[7] Ms. Dipa Dixit Lecturer, Ms. M Kiruthika Assistant Professor, Fr.CRIT, Vashi, "Preprocessing Of Web Logs", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 07, 2010, 2447-2452

[8] Ankit R Kharwar1, Chandni A Naik2, Niyanta K Desai3, 1Assistant Professor, Department of Computer, Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli , 2,3Student of M.Tech Computer Engineering in Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli, "A Complete Pre-Processing Method for Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 10, October 2013)

[9] Surbhi Anand, Surbhi Anand, Department of Computer Science & Engineering, Thapar University, Patiala-147004 (India), "An Efficient Algorithm for Data Cleaning of Log File using File Extensions", International Journal of Computer Applications (0975 – 888) Volume 48– No.8, June 2012

[10] Gopal Pandey, Swati Patel, Vidhu Singhal, Akshay Kansara, "A Process Oriented Perception of Personalization Techniques in Web Mining", International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Volume-1, Issue-2, January 2013

[11] V. Shanmuga Priya1, S. Sakthivel, Department of computer science, Periyar University, TamilNadu, India, "An Implementation Of Web Personalization Using Web Mining Techniques", International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, IJCSMC, Vol. 2, Issue. 6, June 2013, pg.145 – 150

[12] V. Sathiyamoorthi, Department of CSE, Sona College of Technology, Salemi-5, and Dr.Murali Bhaskaran, Principal,Paavai College of Engineering, Paachal, 637018, Tamil Nadu, India, "Data Preprocessing Techniques for Pre-Fetching and Caching of Web Data through Proxy Server", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.11, November 2011

[13] Ramya C, Dr. Shreedhara K S and Kavitha G, M.Tech (Final Year), Professor & Chairman and Lecturer, Dept. of Studies in CS&E, U.B.D.T College of Engineering, Davangere Davangere University, Karnataka, INDIA cramyac@gmail.com and ks_shreedhara@yahoo.com, "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process", International Conference on Communication and Electronics Information (ICCEI 2011)

[14] Abdul Rahaman Wahab Sait, and Dr.T.Meyappan, "Data Preprocessing and Transformation Technique to Generate Pattern from the Web Log", International conference on Computer Science and Information Systems (ICSIS'2014) Oct 17-18, 2014 Dubai (UAE)

[15] Wasvand Chandrama, Prof. P.R.Devale, Prof. Ravindra Murumkar, Department of Information technology, Research scholar of Bharati Vidyapeeth University College of Engineering, Pune, Maharashtra 411046, India., ISSN 2348 – 7968, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, December 2014.

[16] Sheetal A. Raiyani, Shailendra Jain, Dept. of CSE(SS),TIT,Bhopal], "Enhance Preprocessing Technique Distinct User Identification using Web Log Usage data", ISSN:2249-5789, International Journal of Computer Science & Communication Networks,Vol 2(4), 526-530

[17] Michal Munk, Jozef Kapusta, Peter Švec, Constantine the Philosopher University in Nitra, Department of Informatics, Tr. A.Hlinku 1, 949 74 Nitra, Slovakia, "Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor", International Conference on Computational Science, ICCS 2010

[18] Mr. Shivkumar Khosla, Mrs. Varunakshi Bhojane, Department of Computer Engineering, Mumbai University, India, "Capturing Web Log and Performing Preprocessing of the User's Accessing Distance Education System", International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.2, Issue.5, Sep.-Oct. 2012 pp-3128-3130 ISSN: 2249-6645

[19] Doru Tanasa and Brigitte Trousse, AxIS Project Team, INRIA Sophia Antipolis, "Advanced Data Preprocessing for Intersites Web Usage Mining", 1094-7167/04/$20.00 © 2004 IEEE 59, Published by the IEEE Computer Society

[20] Mr. Rahul Mishra, Ms. Abha Choubey, Computer Science & CSVTU India, "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, Volume 2, Issue 9, September 2012

[21] Nanhay Singh, Achin Jain, Ram Shringar Raw , nsingh1973@gmail.com , achin_jain25@yahoo.com , rsrao08@yahoo.in, Ambedkar Institute of Advanced communication Technologies & Research Delhi, India, "COMPARISON ANALYSIS OF WEB USAGE MINING USING PATTERN RECOGNITION TECHNIQUES", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013

[22] Chaitra L Mugali and Asst. Prof. Padma Dandannavar, "WEB LOGS PRE-PROCESSING AND ANALYSIS: A Survey", International Journal of Emerging Technology In Computer Science and Electronics, Volume 14, Issue 2, ISSN: 0976-1353

