II. LITERATURE SURVEY

Ankita Kusmakar, Sadhna Mishra [1] says the Web has evolved by huge development in number of websites and number of accessing users. So in order to provide better service along with enhancing the quality of websites, it has become very important for the website owner to better understand their customers. This is done by mining web access log files to extract interesting patterns. This paper presents an overview of web usage mining, its techniques & also provides a summary of the pattern extraction algorithm used for Web Usage Mining.

Seong Dae Lee, and Hyu Chan Park [2] says consider website structure as weighted directed graph and user navigation path as traversal on it where node represents webpage, edge represents link between Webpages, weight of node represents browsing time of webpage. Then we apply graph mining method.

Data which is used for web usage mining can be collected at three different levels [3].

A. Server Level:
The server stores data regarding request performed by the client. Data can be collected from multiple users on single site.

B. Client Level:
It is the client itself which sends information to a repository regarding the user’s behavior. This is done either with an ad-hoc browsing application or through client side application running standard browsers.

C. Proxy Level:
Information regarding user behavior is stored at proxy side, thus web data is collected from multiple users on several web sites, but only users whose web clients pass through the proxy. In this paper we will cover only the case of web server (HTTP server) data. The information that we have at the beginning is automatically collected by web server and it is stored in access log files, CERN and NCSA specified a common log format (CLF) for every access stored in a log and it is supported by most of the HTTP servers. Following is an example line of access log in common log format

db01.grohe.it[19/sep/2001:03:23:53+0100]
“GET/HTTP/1.0” 200 4096

III. PROPOSED SYSTEM

The proposed system gives many data preparation techniques to clean the data and identify users and sessions. Also client side data is handled to overcome the problem of caching and existence of proxy servers. The system mainly handles websites where users may not be comfortable revealing their identities by logging in. Hence the system will identify users and then identify sessions using two-way
hash structure. Due to the hash structure used in storing user session sequence, backward referencing is done without the entire server pages tree being searched. Hence backward referencing takes much lesser time and also identifies session with higher precision.

IV. STAGES IN OUR PROPOSED SYSTEM
1) Data collection
2) Parse the log by extracting the fields.
3) Store the data in a relational database
4) Merge data from various sources stored in intermediate files.
5) Data cleaning
6) User identification
7) Session Identification
8) Path completion
9) Data Summarization

V. ALGORITHM
1) Start
2) For each log entry
   If request was unsuccessful then move to anomaly Table (check status)
   Else if the request is from web robot then move to Anomaly table (find browsing speed or —robots.txt in filename)
   3) Keep a count of each kind of requests pruned and saved in anomaly table.
   4) Stop.

VI. METHODOLOGY
A. User Identification:
It means identifying each user accessing website and the goal is to mine every user's access characteristic and then make user clustering and provide personalized service to them. It is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers [4]. If login is provided on the website then the user can directly be identified based on the login identification, but in absence of login information, following heuristics are considered.
1) IP Address:
   Each IP address represents one user. Since there might be multiple users with same IP address, only IP address information not reliable.
2) User Agent:
   For more logs, if the IP address is the same, but the agent log shows a change in browser software or operating system, an IP address represents a different user. Hence IP address along with the user agent is matched to identify unique users.

B. Session Identification:
Session is the time between the logged in and logged out and finds the sequence of clickstream to trace the user activity.
Following rules are briefed to identify user session in the project:

1) If there is a new user, there is a new session.
2) In one user session, if the refer page is null, there is a new session.
3) If the time between page requests exceeds a limit of 30 minutes (default timeout for session). It is assumed that the user is starting a new session.

VII. CONCLUSION

The system mainly handles websites where users may not be comfortable revealing their identities by logging in. Hence the system will identify users and then identify sessions using two-way hash structure.

In future the system could be modified so as to recognize if the users in the new batch of log file already exists in the earlier batches of log files processed.

VIII. ACKNOWLEDGEMENT

It gives us great pleasure in presenting this project report titled: “Web log pre-processing for web usage mining”. On this momentous occasion, we wish to express our immense gratitude to the range of people who provided invaluable support in the completion of this project. Their guidance and encouragement has helped in making this project a great success.

We express our gratitude to our project guide Prof. Samira Nigrel, who provided us with all the guidance and encouragement and making the lab available to us at any time. We also would like to deeply express our sincere gratitude to Project coordinators.

We are eager and glad to express our gratitude to the Head of the Information Technology Dept. Prof Neelima Pathak, for her approval of this project. We are also thankful to her for providing us the needed assistance, detailed suggestions and also encouragement to do the project.

We would like to deeply express our sincere gratitude to our respected principal Prof. Dr. Shrikant Kallurkar and the management of Atharva College of Engineering for providing such an ideal atmosphere to build up this project with well-equipped library with all the utmost necessary reference materials and up to date IT Laboratories.

We are extremely thankful to all staff and the management of the college for providing us all the facilities and resources required.

REFERENCES


