Analysis of User Behavior through Web Usage Mining

Suhasini Parvatikar
Department of Computer Engineering
Saraswati College of Engineering, Kharghar

Bharti Joshi, Ph.D.
Department of Computer Engineering
Saraswati College of Engineering, Kharghar

ABSTRACT
Internet has gained a lot of interest of customers. Maintaining good relations with customer is the major applications in web mining. The process of extracting useful information from web logs is called Web Usage Mining. This helps to improve the website performance by analyzing the user's interest and also adds profitability in business The main goal of Web Usage Mining is to study the users' navigation patterns and their use of web resources. Web Usage Mining is the primary focus of this study and we will learn more about the different stages involved in this mining process and with comparative analysis between pattern discovery algorithms i.e Apriori and FP-growth algorithm**.**

Keywords
Apriori Algorithm, FP-growth Algorithm, Log file, Preprocessing, Web Usage Mining, and Web mining.

1. INTRODUCTION
Web mining[3] is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. It[3] is used to understand customer behavior, improve the performance of website to make business profitable. Figure 1 shows that Web mining can be categorized into three different types that are Web usage mining, Web content mining and Web structure mining .

1. **Web usage mining:** In this step, web logs data is collected and interested patterns are discovered and analyzed.
2. **Web content mining**: In this step, text and pictures are mined .
3. **Web structure mining[15]**: The[15] structure of website are mined on basis of hyperlinks and intra-links.

The objective of this paper is to concentrate on the overview of Web usage mining and information present in log file so as to study about users behaviour and improve the performance of website . Figure 2[11] shows the basic process of uses mining process and the applications. The[4] data is collected from the web servers, these web server manage the web access log for all the websites that are managed under the web server. This[4] data is cleaned and managed in a specific format and may be used for different kind of information extraction using different schemes.

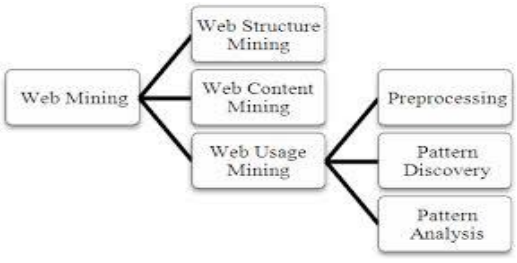

Fig 1: Taxonomy of Web mining

2. LITERATURE REVIEW
Various analysis tools like Web trends, Open Market Web Reporter, Net genesis etc. have been developed to work with the data from the server logs for Web Usage Mining and also help with reporting user's activities. But these tools have limited capabilities in dealing with huge data and in understanding and analyzing the relationship between the data in the server logs.

*Mirghani. A. Eltahir , Anour F.A. Dafa-Alla [1]*introduced this capability of using the data mining techniques to extract information from the server logs with the stages of WUM. It also shows some important aspects like data exploration ,analysis of activity and preferences of users.

Rahul Mishra, Abha choubey [2] discusses about the use of web page accesses from the different server logs to discover the frequent usage by the client and also from the experimental study ,finds some interesting patterns through association rule mining algorithm and compares between pattern mining algorithm i.e Apriori and Fp growth algorithm.

It is very important to collect reliable user usage data and the way one can achieve this is by determining problems like system errors, corrupted and broken links and errors which arise in Web Surfing is discussed by *K. R. Suneetha, Dr. R. Krishnamoorthi* and this obtained results of the study will be used in the further development of website in order to increase its effectiveness.

S. K. Pani, , L. Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Padhi, [16] provides a survey of the pattern extraction algorithms used for web usage mining .In order to discover the mining of traversal patterns, the concept of maximal forward reference was introduced to break down the user session into transactions. The maximal forward reference basically refers to accessing the last page before a call for the previous viewed page is made in that particular user session. The mining algorithms that were developed and introduced in this paper.

R.M. Suresh, R. Padmajavalli [13] discusses the importance of data preprocessing methods and various steps involved in getting the required content effectively. This paper focuses the processes of data preparation and transaction identification which leads to development of the user session file and transaction file.

Web Usage Mining is the primary focus of this study and we will learn more about the different stages involved in this mining process and with comparative analysis between pattern discovery algorithms i.e Apriori and FP-growth algorithm is the objective of this paper.

3. WEB USAGE MINING
3.1 Web Log Data
A Web log file [3] records activity information when a Web user submits a request to a Web Server. Generally,a log file

can be found in three different places: i) Web Servers, ii) Web proxy Servers and iii) Client browsers.

i) *Server-side logs:* These [3] logs generally supply the most complete and accurate usage data.

ii) *Proxy- side logs*: A[3] proxy server takes the HTTP requests from users and passes them to a Web server then returns to users the results passed to them by the Web server.

iii) *Client-side logs:* Participants [3] remotely test a Web site by downloading special software that records Web usage or by modifying the source code of an existing browser.Mostly, HTTP cookies could also be used for this purpose.

Information[1] obtained from log file is explained as follows:

i) *Number of Hits:* This[1] number of times any resource is accessed in a Website. When a web page is uploaded from a server the number of "hits" or "page hits" is equal to the number of files requested.

ii) *Number of Visitors:[1]* It's a user who navigates to website and browses one or more pages.

iii) *Visitor Referring Website:* The[1] referring website gives the information or URL of the website which referred the particular website in consideration.

iv) *Visitor Referral Website*: The[1] referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

v) *Time and Duration:* The[1] time and duration for how long the website was accessed by a particular user.

vi) *Path Analysis:* Path [1]analysis gives the analysis of the path a particular user has followed in accessing contents of a website.

vii) *Visitor IP Address:* This[1] information gives the Internet Protocol (IP) address of the visitors who visited the website in consideration.

viii) *Browser Type:* This[1] gives the information of the type of browser that was used for accessing the website.

ix) *Cookies:* A[1] message given to a web browser by a web server. The browser stores the message in a text file called cookie. The main purpose of cookies is to identify users and possibly prepare customized web pages for them .

x) *Platform*: This[1] information gives the type of operating system used to access the website .

Below is an Example of Common Log Format, type of standardized log format. The syntax can be given as

127.0.0.1 user-identifier frank [10/Oct/2000:11:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326

A "-" in a field indicates missing data.

i. 127.0.0.1 is the IP address of the client (remote host) which made the request to the server.

ii. User-identifier is the RFC 1411 identity of the client.

iii. Frank is the userid of the person requesting the document.

iv. [10/Oct/2000:11:55:36 -0700] is the date, time, and time zone when the server finished processing the request, by default in strftime format

%d/%b/%Y:%H:%M:%S %z.

v. "GET /apache_pb.gif HTTP/1.0" is the request line from the client. The method GET, /apache_pb.gif the resource requested, and HTTP/1.0 the HTTP protocol.

vi. 200 is the HTTP status code returned to the client. 2xx is a successful response, 3xx a redirection, 4xx a client error, and 5xx a server error.

vii. 2326 is the size of the object returned to the client, measured in bytes.

3.2 Phases of Web Usage Mining

Web usage mining is the process of extracting useful information from server logs. Some of the users are interested in only textual data some are in multimedia data. Thus, Web Usage Mining helps to discover interesting usage patterns from Web log data in order to understand and better serve the needs of Web-based applications. Web Usage Mining[10] is application of data mining techniques.

The [10] main source of data for web usage mining consists of textual logs collected by numerous web servers all around the world. There[10] are four stages in web usage mining.

a. Data Collection: users[10][13] log data is collected from various sources like server side, client side, proxy servers and so on.

b. Preprocessing: Performs[10][13] a series of processing of web log file covering data cleaning, user identification, session identification, path completion and transaction identification.

c. Pattern discovery: Application[10][13] of various data mining techniques to processed data like statistical analysis,association,clustering, pattern matching and so on.

d. Pattern analysis: once[10][13] patterns were discovered from web logs, uninteresting rules are filtered out. Analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

All the four stages are depicted through the following figure 2[11].

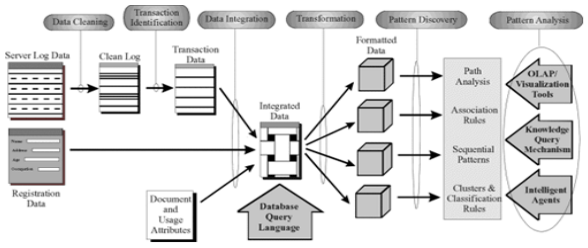

Figure 2: Phases of Web usage mining

3.2.1 Data Collection

Data collection is first step in web usage mining. Data is collected from various sources i.e from server side, proxy side or client side.

a. The server side:
These[11][10] logs usually contain basic information e.g.: name and IP of the remote host,

date and time of the request, the request line exactly as it came from the client, etc. This[11] information is usually represented in standard format.

b. *The Proxy Side:*
Many Internet[11][10] Service Providers (lSPs) give to their customer proxy server services to improve navigation speed through caching. The main difference[11] with the server side is that proxy servers collect data of groups of users accessing groups of web servers.

c. *The Client Side:*
Access data[11][10] can be tracked also on the client side by using JavaScript or applets, or even modified browsers. These techniques avoid the problems of session identification

3.2.2 Data Preprocessing

This is second step in web usage mining. It consists of 3 sub stages i.e. data cleaning, user identification and session identification as shown in figure 3.This is the important step as it transforms[11] data into a format that will be more easily, and efficiently processed for the purpose of the user. The[11] main task of data preprocessing is to select standardized data from the original log files.

I. *Data Cleaning*: In data cleaning phase [12] irrelevant or redundant information like image, video and sound files which could be downloaded without an explicit user request can be removed. Other removal information includes HTTP errors, records created by spiders, crawlers and robots.

II. *User Identification:* Once HTTP [11] log files have been cleaned, next step in the data preprocessing is the identification of users. Different methods for this are 1)by converting IP address to domain name. 2) The web server randomly assigns an ID to web browser while it connects first time to the site. To[3] identify unique users we propose some rules: If there is new IP address , then there is a new user, if the IP address is same but the operating system or browsing software are different, a reasonable assumption is that each different agent type for an IP address represents a different user.

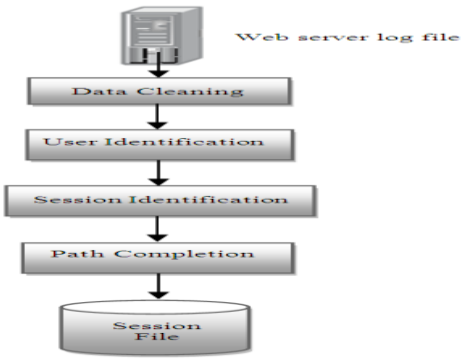

Fig 3: Stages of Data Preprocessing

III. *Session Identification:* To group [11] the activities of a single user from the web log files is called a session. As long as user is connected to the website, it is called the session of that particular user. Most[11] of the time, 30 minutes time-out was taken as a default session time-out. A session[11] is a set of page references from one source site during one logical period. We[11] used following rules to identify user's sessions:

1. If[11] there is a new user there is new session.

2. In[11] one user session, if the referrer page is null, there is a new session.

3. If the[11] time between page requests exceeds a certain limit (30 minutes). It is assumed that user is starting a new session.

IV. *Path completion:* To[11] identify unique user sessions, it is necessary to determine if there are important accesses that are not recorded in the access log. Path Completion refers to[11] inclusion of important page access records that are missing in the access log due to browser and proxy server caching. If[11] a page request is made that is not directly linked to the last page a user requested, the referrer log can be referred to see which page the request came from. If[11] the page is in the user's recent request history, it is assumed that the user backtracked with the "back" button, using cached versions of the pages until a new page was requested.

3.2.3 Pattern Discovery

In the pattern discovery phase[2], Special pattern discovery algorithms applied on raw data which is output of the data processing phase. Typically[5] the first technique applied to the data is *statistical analysis*. With this[5] technique, the type of information extracted is

- Most frequently requested pages;
- Average access time;
- Most common error coded, etc.

Clustering: clustering[16] can be done according to i) Web pages ii) Web page sequences, iii) client IP etc.

Classification: classify users according to their navigational behavior.

Association Rules: discover[7] correlations among pages accessed together by a client for example, thirty percent of department page viewers will enter the cs-dept pages.

3.2.4 Pattern Analysis

This [12] is the final stage of Web Usage Analysis. The goal of this process is to extract the interesting patterns from the output of the pattern discovery process by eliminating the irrelative patterns. Knowledge query mechanism [10] such as SQL is the most common method of pattern analysis. Another[10] method is to load usage data into a data cube in order to perform OLAP operations.

4. ASSOCIATION RULE

Association rule[2] is used to find out the items which are frequently used together. The Presence of one set of items in a transaction implies other set of items. The terms[2] used in these rule are

a. Support: The support[2] of an association rule X implies Y is the percentage of transaction in the database that consists of X U Y.

b. Confidence: The confidence[2] for an association rule X implies Y is the ratio of the number of transaction

that contains X U Y to the number of transaction that contains X.

c. Large Item Set: A large[2] item set is an item set whose number of occurrences is above a threshold or support.

The task of association rule mining is to find correlation relationships among different data attributes in a large set of data items, and this has gained lot of attention since its introduction. Such relationships observed between data attributes are called association rules.

5. COMPARISON BETWEEN APRIORI AND FP-GROWTH ALGORITHM

Apriori algorithm[2] is an algorithm for frequent item set mining and association rule learning .The FP-Growth Algorithm[2] is an alternative way to find frequent itemsets without using candidate generations.

Comparison table:

PARAMETERS	APRIORI ALGORITHM	FP-GROWTH ALGORITHM
Memory Required	Large due to large candidate sets	Less as no candidate sets
No of Visits to DB	Multiple times for candidate sets	Twice only
Techniques used	Use apriori property and join and prune property	It constructs conditional frequent pattern tree and conditional pattern base from database which satisfy minimum support.
Time required for Execution	More due to candidate set production	Small compared with Apriori.

6. CONCLUSION

The main goal of this study was to find the idea of obtaining as much information as possible about the end user from the logs is a promising and less explored form of web log analysis. There are many ways of gaining personal information from the user when a person visits a particular website with the help of cookies, registration forms, java scripts on the client and server ends. But the idea of gaining information from the user IP and then developing some feature sets from the data is novel. Different pattern extraction algorithm can be used to find the frequent pattern and the results of which can be used to structure the website and information that it consists of according to the personal preferences of the respective user category.

7. REFERENCES

[1] Mirghani. A. Eltahir , Anour F.A. Dafa-Alla "Extracting Knowledge from Web Server Logs Using Web Usage Mining" Blue Nile University ,Dmazin, Sudan,IEEE 2011

[2] Rahul Mishra, Abha choubey "Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data." (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (4) , 2012

[3] K. R. Suneetha, Dr. R. Krishnamoorthi, " Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[4] Bhaiyalal Birla, Sachin Patel , " An Implementation on Web Log Mining", IT & PCST, Indore, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014.

[5] Preeti Sharma & Sanjay Kumar, "An Approach for Customer Behavior Analysis using Web Mining,NIT Raipur, Raipur, Chattishgarh, International Journal of Internet Computing (IJIC), ISSN No: 2231 – 6965, Volume-1, Issue-2, 2011.

[6] Aditi Shrivastava, Nitin Shukla, " Extracting Knowledge from User Access Logs", Shri Ram Institute of Technology, Jabalpur, International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012

[7] Kobra Etminani, Mohammad-R. Akbarzadeh-T, Noorali Raeeji Yanehsari , "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method", International Journal of Internet Computing (IJIC), 2009.

[8] S. K. Pani, , L. Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Padhi, "Web Usage Mining: A Survey on Pattern Extraction from Web Logs" International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.

[9] Mr. Rahul Mishra ,Ms. Abha Choubey , " 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.

[10] V.Chitraa,Dr. Antony Selvdoss Davamani, "A Survey on Preprocessing Methods for Web Usage Data", *(IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.*

[11] Vijayashri Losarwar, Dr. Madhuri Joshi , "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems , 2012 .

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

[13] R.M. Suresh, R. Padmajavalli, "An Overview of Data Preprocessing in Data and Web Usage Mining", RMK Engineering College, Kavaraipettai, IEEE 2006.

[14] P.Nithya, Dr. P.Sumathi , "An Enhanced Pre-Processing Technique for Web Log Mining by Removing Web Robots", Tamilnadu ,IEEE 2012.

[15] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood, "Web Usage Mining: A Survey on Preprocessing of Web Log File" , Center of Research in Data Engineering (CORDE) ,IEEE 2006.

[16] S. K. Pani, , L. Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Padhi, "Web Usage Mining: A Survey on Pattern Extraction from Web Logs", International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.