IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Comprehensive Overview On Web Usage Mining Its Task & Techniques

Sonam Singh Gurjar[1*], Khushboo Agrawal[2]

[1,2]Dept. of Computer Science and Technology, MITS, Gwalior, India

Corresponding Author: sonamsingh2450@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i5.590599 | Available online at: www.ijcseonline.org

Accepted: 10/May/2019, Published: 31/May/2019

***Abstract* —**Internet users in the world increasing rapidly. At the present time, the best way of conveying information is the World Wide Web. There are so many websites for learning, shopping, selling, businesses and many more. The expansion of Internet usage will result in increasing web data speedily. To exploit the information of internet usage, it becomes necessary to extract the access behavior of the users.Web usage mining is one of such Data mining technique used for mining, web access log.These access logs are saved on the web server.Access log is the records of all the user, requests for a particular file from a website.Web usage mining will help in improving the design of the website and the personalization of the content. This paper gives the comparative study of web usage mining, it also summarizes the web usage mining approach like pre-processing, pattern discovery, pattern analysis, visualization. This survey listed various research work done by the researcher. It delivers numerous techniques and algorithms used in web usage mining.

***Keywords*—** *server log, access log, web usage mining, pre-processing, user identification, session identification, clustering, classification, pattern discovery & analysis.*

I. INTRODUCTION

Data mining is considered a technique for finding useful and fascinating patterns from the data. This data can be stored in any database, information repositories and data warehouses in the form of text, image, log files, numerical data, etc [1]. Web mining is the interdisciplinary field of Data Mining which includes techniques of machine learning, artificial intelligence, statistics databases, visualization. The task of web mining contains classification, clustering, and association rule mining and pattern discover& analysis [2]. Classification and Clustering are supervised and unsupervised learning respectively.

The main process of data mining includes:

- **Pre-processing:** Usage data we got from different resources are not suitable for mining because of data contains noise and abnormalities. There are irregularities in data, sometimes data are too large to mine. To overcome all these efficient pre-processing techniques must be followed, which helps us to get better results.
- **Data–mining:** Various data mining techniques are then implemented on pre-processed data for taking out useful knowledge.
- **Pattern Discovery:** Not all patterns are useful; this is the task of determining effective patterns as according to the application.

In data mining association, we have basically three categories of mining named as text mining, data mining, web mining. Data Mining mostly agrees with the structured data organized within the database. It primarily focuses on data dependent activity. One can quickly deploy defined algorithms for better results. It helps in predicting outcomes from large databases. Text Mining primarily handles unstructured data or text. With the growth of social networking, it becomes necessary to mine text data for doing sentiment analysis and predict information assets that are strategic. Web Mining hooks with unstructured data. Web data mining requires innovative use of data mining or text mining techniques and approaches that are idiosyncratic. It finds relevant information from the Web by using structural links, contents or some log statistics [4].

With the growth of worldwide network internet access becomes in hand for everyone. The amount of web data is huge. One can find information on every topic. All commercial websites allow users to visit their site and perform useful operations.

We have powerful searching tools to find information quickly and specifically on the web. There is a requirement

for decreasing traffic load and projecting the website in a way that is suitable for various users. To achieve these requirements web providers desires to discover a technique to envisage user's behavior and personalization information [1].

Web usage mining is one of data mining technique which is used for determining interesting patterns from web log files. Extraction of likely usage patterns, knowledge and information commencing web hyperlink structure, page content, and web user's data is called as web mining [3]. Web mining is an important part of data mining where it mainly handles information like structure, content, log data. On the basis of this information, Web Mining is basically separated into three parts i.e.

- WEB CONTENT MINING
- WEB STRUCTURE MINING
- WEB USAGE MINING

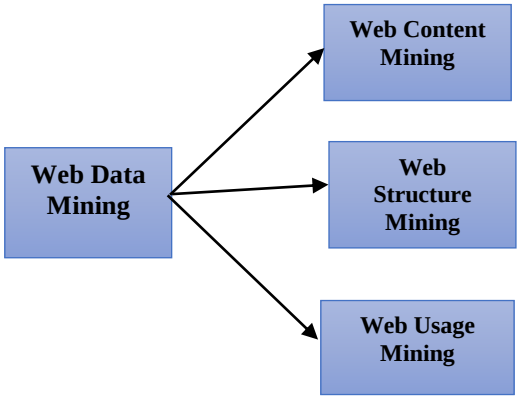

Figure 1. Web Data Mining Structure

Web Content Mining It extracts useful patterns and mining knowledge from those contents collected from the web. The tasks of content mining are somewhere close to our traditional determining. The most important strategy of content mining on the web is data retrieval out of the web content. For example, one can easily cluster and classify web contents as a rendering of the topic.

Web Structure Mining: In this approach of web mining valuable information about hyperlink structures is extracted. The Hyperlink is the operator module that links a web page to an alternative site. It uses graph mining and network mining concept and method for examining connections and node structure. For example, in using web links one can determine significant web page, that is a primary technology employed by many search engines.

Web Usage Mining: It ascribes detection of user access patterns through usage logs on the web, that takes down user's usage data stored on the web. The topmost issue in web usage mining is the pre-processing of usage logs for producing applicable data for mining [2]. It is the task of analyzing and uncovering the actions of web clients when they are surfing and navigating on the Web. Usage mining on the web is used to amplify the feature of e-commerce website services, for personalizing the websites. By using various mining techniques, we can get a better web structure and web server performance [5]. This survey paper mainly concentrates on web usage mining architecture, techniques used for data gathering, data pre-processing, data modeling, pattern discovery and pattern analysis. In this survey, we also discuss web usage mining applications and its future scope.

1.1 System Architecture of Web Usage Mining Process

There is a wide range of infestation and implementations exercised by Web Usage Mining. Web mining uncovers the user access pattern through mining log files and linked data of the particular web portal. The principal outset of data for usage mining is the web log files on the server. The main process in the web usage mining architecture includes [6].

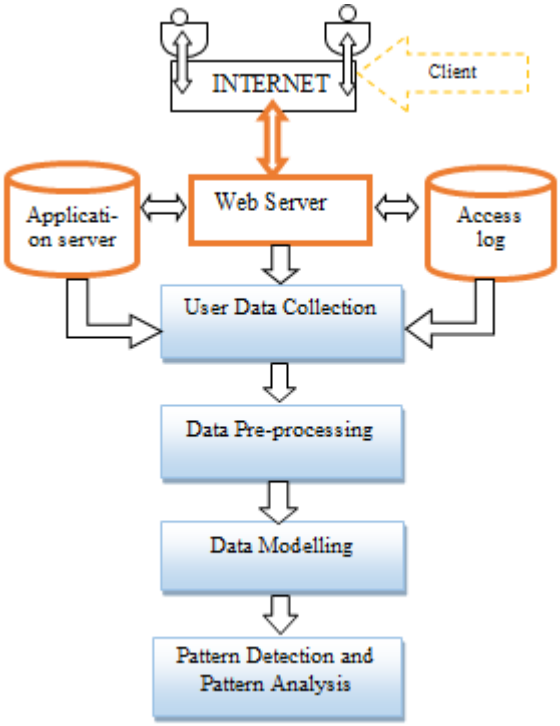

Figure 2. System Architecture of Web Usage Mining

- **Usage Data Collection:** This is an important task of web usage mining. Usage data which record user activities in the form of log files are gathered for mining. The primary sources of data in web usage mining are application servers.

- **Data pre-processing:** Data pre-processing includes data fusion, page view identification, and session identification and synchronization. It involves removing inappropriate references and cleaning some type of data fields (HTTP protocol used, number of bytes transferred).
- **Data Modelling:** from the previous process, we have a class for the page views: P= $\{p_1, p_2....p_n\}$and a class for user transactions: T= $\{t_1, t_2.....T_n\}$ in there every transaction t_i contains a subclass for P. For many mining algorithms sequential ordering of the transaction is not expedient, to overcome this, transactions can be represented an n-length vector.

	Page 1	Page 2	Page 3	Page 4	Page 5
User A					
User B					
User C					
User D					
User E					

Figure 3. user page views matrix for transaction matrix

- **Pattern Extraction and Analysis:** In this process of web usage mining we discover useful patterns from the server log. A number of the pattern discovery methods are statistical analysis, clustering, classification, sequential pattern mining, and association rule mining.

1.2 Usage Data Collection

The footstep of web usage mining is to collect data that are suitable for mining from various sources. Data are imparted on the web mainly by testing performed by human or usage data on the web server log.[6] When user request to a web server all its activities are recorded. Three major sources where log files are recorded as a server-side log, client-side log, and intermediary data.

- Server-side log: user request is recorded in the Web Server as weblogs. These logs contain private, subtle information. It includes log files, unambiguous user input, cookies. The main resource for web usage mining is the server log which contains various types of logs. The most popular weblogs are NCSA Common Log, Microsoft IIS Log, etc. The different web server supports different log format.

  Common Log Format (CLF): It keeps track of user request which occurs on a website in sequential order. This format holds the information like data related to clients IP address, access time and date, the status returned by the server, URL, bytes transferred. The description of sample log entries is given below:

1) Remote host: This is the IP address of the remote user that made the request.
2) Base URL: URL of user request.
3) Timestamp: It gives the client's a visit date and time.
4) Request method: Request operation is referred to as HTTP. Request methods are cacheable, idempotent, harmless. GET, HEAD, POST is the request methods used by the client. GET method requests should only retrieve data. The POST method is used to acquiesce an entity to specific resource instigating a variation in the state. HEAD is the method which asks for a response alike to that of GET request, except response body.
5) File: Specific file request during the page search.
6) Protocol: Protocol used by the client.
7) Status Code: It indicates the success or failure of the HTTP request. It consists of three digits. For example, 200 is for success, 500 for Internal Server Error, 404 is for File Not Found
8) Bytes Transferred: Number of bytes transferred to the user.
9) Referrer: This is the URL for the mentioned

```
109.169.248.247 [12/Dec/2015:18:25:11 GET /administrator/ HTTP/1.1 200 4263
Mozilla/5.0 (Windows NT 6.0; rv:34.0) Gecko/20100101 Firefox/34.0
46.72.177.4 [12/Dec/2015:18:31:08 POST /administrator/index.php HTTP/1.1 200 4494
Mozilla/5.0 (Windows NT 6.0; rv:34.0) Gecko/20100101 Firefox/34.0
```

Figure 4. Sample Web Server Log Entries

server.
10) User-Agent: This specifies the software or operating system used by the client for accessing the site.[7]

Extended Log Format: It contains a sequence of lines which contains ASCII characters as provided by the World Wide Web Consortium W3C. The following directives are defined: Version, Fields, Software, StartDate, End Date, Remark. Extended Log Format supported by the Netscape web server, W3SV, and Apache. For example, #Version:1.0 #Date: 12-Jan-1996 00:00:00 #Fields: time cs-method cs-url 00:34:23 GET /foo/bar.html 12:21:16 GET/foo/bar.html 12:45:52 GET/foo/bar.html 12:57:34 GET/foo/bar.html

Unambiguous user input: this is the type of data obtained from the registration forms submitted by

the user. Such data are often erroneous or inadequate.

Cookies: These are the text files which automatically generates on the client's browser. Cookies data can be used for user next visit. The User ID is sent to the web server together with the request made by the user.[8]

- Client-side log: These are the log data stored on the client browser. Java applets and Java scripts are the remote agents used to gather user browsing info.
- Intermediary data: This type of data is created between the web server and browsers. Proxy servers are the servers where all the intermediary data stored. It records request and response of web pages from the web server. Proxy caching will reduce the loading time of any web page.[8]

1.3 Data Pre-processing

There are a huge number of requests stored on the web server, not all log data are reliable to perform usage mining. The main reason for unreliability is IP address delusion or web caching. Logs are usually stored in the text format will be comprised of immense data having incomplete and undesirable data too.

This is a process in which raw server logs stored in the various data sources are transformed into a suitable data file for performing mining techniques.

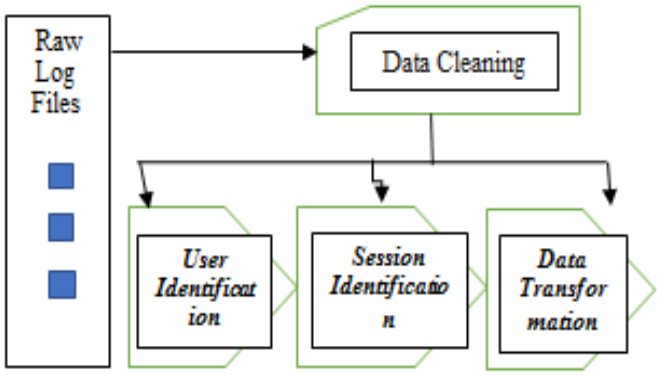

Figure 5. Data pre-processing phase

Various analytical techniques elaborated in data pre-processing are:

A) Data Cleaning:

In this method, redundant log entries and inappropriate data are removed from the log files. Various useless data and missing values are eliminated. The user request for a page which results in several unnecessary log entries, these additional requests are inappropriate for usage mining so needs to be removed. URLs having file extensions like .gif,.jpg,.css,.png,.js, etc are removed.

Robot/crawler entries: Web robot (WR) is the software tool which extracts Website content through scanning. Robot and Spider inevitably track all the hyperlinks from a web page. For efficient mining, we should remove all the user agents having bot, spider, Yandex, crawler. The host having requested page 'robot.txt' are removed.

Erroneous status codes: All HTTP Status code has entries lower than 200 and more than 299 are eliminated. As they show failed status for any requested web page.

We should only keep imperative fields like IP address, Date and Time, URL requested, Time is taken a user agent which will simplify the mining task.[9][11]

B) User Identification:

After data cleaning user identification is the next important task in the usage log pre-processing in this task, we identify the unique number of visitors. This can be done by discovering visitors having a unique IP address. If the IP address is different from the previous one is considered to be a new user. Algorithm for user identification works in the following way: successive accesses of two IP address are compared.[10]

1. Begin
2. For each N record of weblog table
3. Repeat for each IP address
4. If an IP address is in IP table and user_agent is same then
5. Assign old user id
6. Else
7. Assign new user id
8. Increment user id
9. End if
10. End

C) New Session Identification:

Session identification aims to identify different sessions used by the user. It separates the page access of each user into different sessions. Each weblog entry is taken as a session. Session identification defines how many times the user has accessed any webpage, for this a time out mechanism is used. The mechanism used in this is:[11]

=> If there is a new user then there is a new session.

=> Browsing period of a page accessed by the user is calculated by concluding difference within two successive entries.

=> If the browsing time exceeds a limit defined by the browser (generally it will be 30 or 25 minutes), it is believed that there is a new session.

D) Data Transformation:

Data transformation is one of the Data Pre-processing steps in which data are transformed into the applicable format that is relevant to perform various data mining techniques. Usually, raw data have been in text format and is not a suitable format to perform mining

algorithms. Text data are converted into a tabular form for better results.

1.4 Pattern Discovery

In Pattern Discovery phase various techniques are applied to extract knowledge and perceive exciting patterns. Several mining algorithms are used depending on the requirements of the predictors. Data mining approaches are used for discovering unrevealed and interesting patterns from the pre-processed web log data. The main mining techniques employed in web usage mining are:[12]

a) Path analysis:

Path analysis helps in identifying frequent traversal of web pages. It is done by using Graph models. Web pages that are frequently used are represented by the nodes in the tree and links between web sites are represented by edges of that tree. Using tree representation one can determine in-depth knowledge about user navigation patterns. Through path analysis, it becomes easy to know what path do user traverse previously before going to a particular URL.

b) Association rule:

Association rule mining will mention the web pages which are accessed together in a Single server session. In this mining technique, the correlation between the pages that are referenced at the same time is identified. It will also refer frequently accessed pages by the web user. Usually, association discovery techniques are based on the Apriori Algorithm. Along with Apriori, many other algorithms are also used: Eclat, FP Growth, Partition algorithm.

c) Sequential pattern discovery:

Predictions regarding visiting patterns and valuable user trends are discovered using sequential pattern discovery. It helps in discovering inter transaction patterns. A set of Web pages followed by another page in the time stamp order will discover through sequential pattern discovery.

d) Classification:

The classification works on surveying elementary data items within multiple predetermined classes. Developing the profile of users, which relates to any specific class. This needs a supervised learning method such as naïve Bayes classifiers, K nearest neighbor classifier, Decision tree, Support vector machine. These classification methods play a key role in analyzing web applications that will represent the users agreeing to several predefined metrics.

e) Clustering:

Clustering describes techniques for the alliancing set of items into various clusters having similar features. Basic cluster types include usage clusters and page clusters. Usage clusters are the clusters which include web pages that are grouped together, and page cluster includes web pages grouped with respect to web page content.

1.5 Pattern Analysis

This is the final stage of the usage mining operation. Foremost need for pattern analysis comes down to exclude undesirable patterns, models that we have from the pattern discovery phase. Various visualization techniques are used for pattern analysis. Knowledge discovery techniques and OLAP techniques are some basic methods used for discovering hidden patterns from the usage data.

Below figure shows the general statistics used for visualization of daily entry web pages from usage log.

Below table shows various techniques used in web usage mining operations.

Table I. Various steps in web usage mining and their associated techniques

Pre-processing	**Pattern discovery**	**Pattern analysis**
Data cleaning	clustering	Decision making
Session identification	classification	visualization
Data integration	Association rule mining	Online analytic process (OLAP)
Data transformation	Sequential pattern mining	Query mechanism

Figure 4 shows at what stage web mining techniques are applied and how they work.

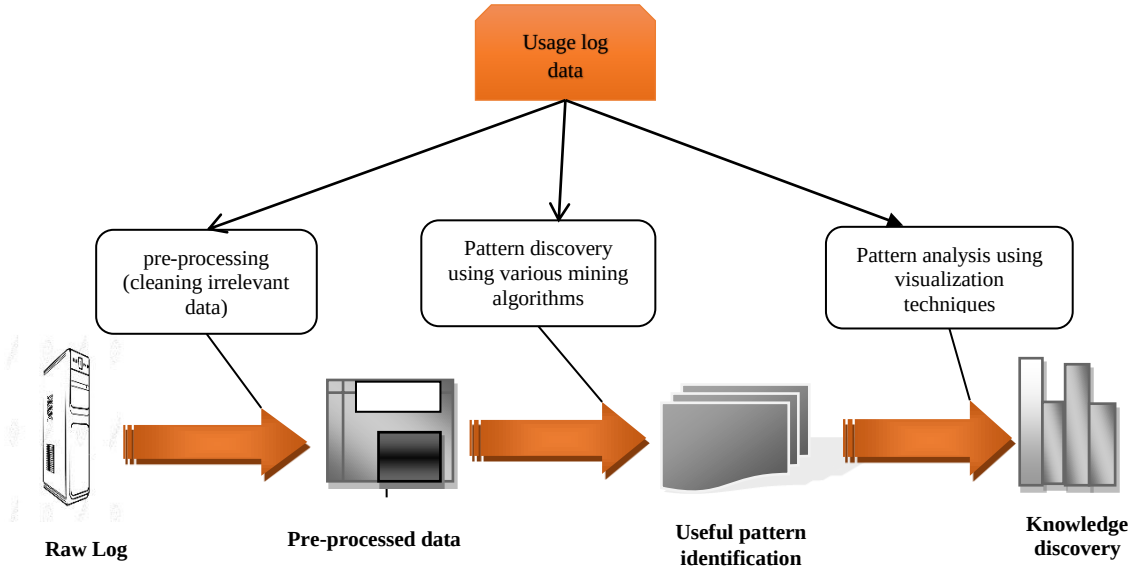

Figure 4. Phases of Web usage mining

TABLE II. WEB USAGE MINING PROCESS ANALOGY

Web usage mining techniques	**Data Pre-processing**	**Pattern Discovery**	**Pattern Analysis**
Data source	Weblogs, Website data, Users login data, Cache, Cookies, etc.	Cleaned log data from the pre-processing phase, Session logs, Transaction logs, etc.	Data from Pattern discovery, Multidimensional database
Process applied	Data cleaning, Session identification, User identification, Data transformation.	Classification, Clustering,	Drill down/up, Roll up
Advantages	Convert raw log data into understandable formats	A Selection of useful information from discovering pattern observation	Extraneous and inappropriate rules are separated
Algorithm	Attribute discretization/Elimination Apriori, FP Growth.	K means, Fuzzy C means, Decision tree, Naïve Bayes, SVM, etc.	SQL, OLAP

This survey paper is prepared in the subsequent way. Section I contains the detail introduction of web usage mining. Section I describes the main process and categories of data mining. Later on, it describes system architecture with step by step process description of web usage mining. This paper gives an in-depth understanding of various pre-processing methods, pattern discovery and pattern analysis techniques respectively. Section II presents related research work. Section III concludes future work associated with web usage mining. It describes, what efforts the researcher must put for getting better efficiency in usage mining. Section IV presents the conclusion of the whole paper. The main purpose of the paper and outcome which we get from it. Section V provides acknowledgment.

II. RELATED WORK

S. Padmajaet al. [16] Analyse User's Behaviors and Growth Factors by using enhanced K means algorithmic rule. In which K means computation enhanced. According to author K means algorithm, the program chooses some data items from the initial centroids. Individual data points, then allotted in the cluster having the neighboring centroid. New cluster centroids are then calculated and the process will continue till the convergence standards are encountered. This improved K means clustering algorithm reduces the overall squares of the gap.

Reeny Zackarias [19] uses a session identification technique predicting similarity in user behavior cleaned log file used for the identification process. This research paper identifies a number of sessions for each user and individual users having the same session.

Jiaoling Du et al. [20] Presented advanced DC Apriori algorithm for reorganizing data warehouse through improved relations of recurrent group of items, it is essential to produce K items that comes frequently to join a 1 frequent group of items by using k 1 frequent itemsets that will reduce several associates and receive recurrent set of items through one prune operation. The process completes by evading unacceptable candidate sets and reduces database scan. They proved that DC Apriori algorithm advantage based on a matrix. They also proved that Less running time of this algorithm with the same result.

Rajashree Shettar [13] Uses a sequential tree algorithm and Generalized sequential pattern (GSM) algorithm for discovering frequent sequence in the log data. Further, various comparative studies of algorithms are proposed. It finds sequential patterns from forming a tree. This algorithm detects running time and number of patterns generated.
Various mining rules for sequential pattern analysis has been projected such as Rule Growth [14] and ERMiner [15], that one embraces a vertical approach and pattern growth in order to conclude rules.

Viswanathan K et al. [17] gave Performance Comparison of C4.5 and K NN Classification techniques. The accuracy of classification algorithm validates by using terms like error rate and computation time. This research discusses various classification algorithms and compared dataset performances. It states that C4.5 is the best classification algorithm which gives enhanced results.

Ketan D. Patel et al. [18] purpose algorithm for pre-processing web log data. Pre-processing algorithms are applied to weblog data of two websites, in this research paper pre-processing task are performed by implementing algorithms for data cleaning, unique visitor's identification. The pre-processing task is required to discover interesting patterns of end users.

V Chitraa et al [21] Calculated effectiveness of navigation patterns using specifications like frequency, downloads, and utility. The techniques were applied after effective pre-processing on all datasets. They have taken constraints like Rand Index measure and Sum of Squared Error. Feature selection was done with the help of Index Components Analysis (ICA). Clustering of navigation patterns was performed by using Bolzwano Weierstrass Theorem (BWFCM), this optimized clustering proves the significance of it. After clustering Classification was done by Vector Machine.

P. Sukumar et, al [22] Effectively classified various data mining process and examined limitations of the algorithms. The in-depth explanation was given for various revised algorithms. Pre-processing of the raw log was done and then cleaned the log file used for identifying the effectiveness of data mining techniques. Unique user identification and session identification are obtained from the cleaned log data.

Swapnil S Patil and H P Khandagale [23] presented an effective way of retrieving useful information from the huge amount of data kept on the web host server. This paper uses web mining techniques for improving web navigation usability. It also delivers a method for updating web links by using sequential pattern mining. It gives a useful method that will help users as well as web developers.

S Sharma and S S Lodhi [24] provide a method for discovering knowledge from a log file which is helpful for extracting information, finding useful patterns. The researcher used the decision tree algorithm which is one of the effective mining methods. Performance evaluation was done by using N fold cross-validation technique and then the classification of log data was done with the decision tree algorithm with some modification.

Table II. Associated work in web usage mining

Paper	Author	Year	Title	Method Used
1	Shilin He [25]	2018	Identifying Impactful Service System Problem via Log Analysis	Employment of Log3C, new clustering-based method
2	Sonia Sharma [26]	2017	Customer Behavior Analysis using Web Usage Mining	Prediction of user behavior on a website data by using pattern discovery and pattern analysis
3	B. Rajeswari [31]	2018	Web Page Prediction Using Web Mining	Behavior analysis is done by using K means clustering, c4.5, SVM, AdaBoostM1, Rule Part
4	Jayanti Mehra [30]	2018	An Efficient method for Web Log Pre-processing and Page Access Frequency using Web Usage Mining	Purposes Algorithm for data cleaning also gives an approach for counting page access frequency.
5	V. Chitraa []	2015	Clustering of Navigation Pattern Using Bolzano Weierstrass Theorem	Calculated Effectiveness of navigation pattern, Feature selection was done by using ICA and BWFCM
6	Ketan D. Patel [18]	2017	Pre-processing on web server log data for web usage pattern discovery	Gives Effective Pre-processing technique by exercising on two websites data
7	ReenyZackarias [19]	2017	Predicting User with Similar Behavior Through Sessions	Session Identification Technique was employed for detecting similarity in user behavior
8	MadihahMohd Saudi [27]	2017	An Efficient Data Transformation Technique for Web Log	New Classification and transformation approach used for getting enhance results. IBK is Used for better result
9	SS Lodhi [24]	2016	Development of Decision Tree Algorithm for Mining Web Data Stream	The Decision tree algorithm was used for efficient mining
10	Tawfiq A. AL– ASDI [28]	2016	An Efficient Web Usage Mining Algorithm Based on Log File Data	Log file Analysis using K means clustering algorithm
11	Arjun Ram Meghwal [29]	2016	Identifying System Error through Web Server Log File in Web Log mining.	Gives a methodology for discovering system error through web server log by using weblog expert tool
12	Aanum Shaikh [32]	2015	Web Usage Mining Using Apriori and FP Growth Algorithm	FP Growth and Apriori
13	Chitra L Mugali [12]	2015	Analysis of Web Server Logs and Pre-processing and	For pre-processing data cleaning, session identification data integration and data transformation were done, further analysis completed with the help of k means clustering, apriori, FP growth

III. FUTURE WORK

We have a large quantity of web usage data on the server. This will create the need for research and techniques associated with web usage mining. In the future, the researcher must explore techniques and algorithms that will help in analysing and predicting user behavior effectively.

Future work must focus on finding effective pre-processing methods as pre-processing is a crucial step in web usage mining. Due to massive web user data, there is a need for enhancing web mining results. A study should be done for increasing efficiency of various usage mining algorithms.

IV. CONCLUSIONS

This paper introduced basic techniques and processes that will consider useful for web usage mining. This research paper enumerated various usage mining approaches for pre-processing and analysing usage data. The main purpose of web usage mining is to extract hidden knowledge from the user's data, to amplify their behavior and finding interesting patterns. Which can further be used to solve many real-world problems like a product recommendation, examine customer behavior, webpage enhancement? This paper specifies pre-processing steps for raw log files. Cleaned log files are used for extracting useful statistics. Unique visitor identification, session identification and data transformation discussed eventually. In pattern discovery phase various clustering and classification, association techniques have been described. This paper gives a survey of recent web usage mining system. Primary operations of web usage mining include data preparation, pattern discovery, pattern analysis, visualization. Each and Every approach has its own pros and cons, however, imperfections can be rectified by advance research.

ACKNOWLEDGMENT

This work has been affirmed by Faculty and Department of Computer Science & Technology of Madhav Institute of Technology (MITS) Gwalior.

REFERENCES

[1] Daniel T. Larose, Discovering knowledge in data: An Introduction to Data Mining, USA: A John Wiley & Sons, INC, publication, 2005.

[2] Bing Liu, Web data mining: Exploring Hyperlinks, Contents, and usage data, German: Springer-Verlag Berlin Heidelberg, pp 527-540, 2007, ISBN 978-3-642-19459-7.

[3] R. Kosala and H. Blockeel, Web mining research: A survey, ACM SIGKDD Explore. 2 (2000) 1–15

[4] Qingyu Zhang and Richards S. Segall, International Journal of Information Technology & Decision-Making Vol. 7, No. 4 (2008) 683–720

[5] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," ACM Trans. Inter. Tech., Vol. 3, No. 1, pp. 1-27, 2003

[6] B.Lalithadevi, A.Merry Ida, A New Approach For Improving World Wide Web Techniques in Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, volume 3,issue1, January 2013

[7] M. Aldekhail, Application and Significance of Web Usage Mining in the 21st Century: A Literature Review, International Journal of Computer Theory and Engineering, Vol. 8, No. 1, February 2016

[8] Murat Ali Bayir, Ismail Hakki Toroslu, Ahmet Cosar and Guven Fidan "Discovering more accurate Frequent Web Usage Patterns," arXiv0804.1409v1, 2008

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

[10] Mr. Shivkumar Khosla, Mrs. Varunakshi Bhojane, Department of Computer Engineering, Mumbai University, India, "Capturing Web Log and Performing Pre-processing 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

[11] V. Chitraa, Dr. Antony Selvadoss Thanamani, A Novel Technique for Session Identification in Web Usage Mining Pre-processing, International Journal of Computer Application (0975 8887) Volume 34 No. 9, November 2011.

[12] Chaitra L Mugali, Pre-Processing and Analysis of Web Server Logs, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, Issue 8, Volume 2 (August 2015)

RAJASHREE SHETTAR ISSN: 2250–3676 [IJESAT] INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY

RAJASHREE SHETTAR ISSN: 2250–3676 [IJESAT] INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY

[13] Rajashree shettar, sequential pattern mining from web log data, IJESAT, ISSN:2250–3676, Volume-2, Issue-2, 204 – 208

[14] P. Fournier-Vige, "Mining partially-ordered sequential rules common to multiple sequences," IEEE Transactions on Knowledge and Data Engineering, vol. 27(8), pp. 2203–2216, 2015.

[15] P. Fournier-Viger, T. Gueniche, et al, "ERMiner: sequential rule mining using equivalence classes," The International Symposium on Intelligent Data Analysis, pp. 108–119, 2014.

[16] S. Padmaja et al., International Journal of Engineering and Technology (IJET), Vol 8 No 1 Feb-Mar 2016

[17] Viswanathan K, Mayilvahanan K, and R. Christy Pushpaleela, "Performance Comparison of SVM and C4.5 Algorithms for Heart Disease in Diabetic", International Journal of Control Theory and Applications, ISSN: 0974-5572, Volume 10, Number 25, 2017.

[18] Ketan D. Patel, "Pre-processing on web server log data for web usage pattern discovery", International Journal of Computer Applications (0975 – 8887) Volume 165 – No.10, May 2017

[19] Reeny Zackarias, "Predicting Users with Similar Behaviour Through Session", International Journal of Advanced Engineering and Research Development (IJAERD) Volume 4, Issue 3, March -2017, e-ISSN: 2348 – 4470

[20] Jiaoling Du, Xiangqi Zhang, Hongmei Zhang and Lei Chen, "Research and Improvement of Apriori Algorithm", IEEESixth International Conference on Science and Technology, pp.117-121,2016.

[21] V. Chitraa and Antony Selvadoss Thanamani, "Clustering of Navigation Patterns using Bolzwano_WeierstrassTheorem", Indian Journal of Science and Technology,Vol8(12),69283, June 2015.PP1-9

[22] P. Sukumar, "Review on Modern Data Pre-processing Techniques in Web Usage Mining (WUM)," International Conference on Computational Systems and Information Systems for Sustainable Solutions,978-1-50901022-6/16/IEEE(2016).

[23] S S Patil and HP Khandagale, "Enhancing Web Navigation Usability Using Web Usage Mining Techniques", International Research Journal of Engineering and Technology IRJET, vol 4 6, June 2016.

[24]S Sharma and S S Lodhi, "Development of Decision Tree Algorithm for Mining Web Data Stream", International Journal of Computer Applications, March 2016.

[25] Shlin He, Qingwei Lin, et al, "Identifying Impactful Service System Problems Via Log Analysis", ESE/FSE'18, November 4–9,2018, lake-Buena-Vista, Florida, USA.
[26] Sonia Sharma et al, "Customer Behaviour Analysis using Web Usage Mining", International Journal of Scientific Research in Computer Science and Engineering, vol 5, issue 6, pp4750, December (2017).
[27] Madihah Mohd Saudi, et al," An Efficient Data Transformation Technique for Web Log", WCE 2017, July 5–7,2017, London, UK.
[28] TAWFIQ A. AL-ASDI, et al, "An Efficient Web Usage Mining Algorithm Based on Log File Data", Journal of Theoretical and Applied Information Technology,31 October 2016 vol 92 No 2, ISSN:1992 – 8645.
[29] Arjun Ram Meghwal and Dr. Arvind K Sharma," Identifying System Error through Web Server Log File in Web Log Mining", International Journal of Computer Science And Technology,Vol.7, ISSN 1, Jan–March 2016
[30] Jayanti Mehra and Dr. R S Thakur, "An Efficient method for Web Log Pre-processing and Page Access Frequency using Web Usage Mining", International Journal of Applied Engineering Research ISSN 0973–4562 Vol–13,November–2(2018),pp1227–1232.
[31] B. Rajeshwari, "Web Page Prediction Using Web Mining", IRJET, Vol:5, Issue 5, May 2018, e–ISSN:2395–0056.
[32] Aanum Shaikh, "Web Usage Mining Using Apriori and FP Growth Algorithm", International Journal of Computer Science and Information Technology, Vol– 6, pp 354–357,2015

Authors Profile

Sonam Singh Gurjar, received her Bachelor of Engineering (BE) in Computer Science and Technology. Currently Pursuing her MTech From Madhav Institute Of Technology and Science (MITS), Gwalior, MP. Her area of interest includes Data Mining, networking, image processing.

Khushboo Agarwal, received her Bachelor degree (B.E.) in Information Technology in 2005, Master degree (MTech) in 2015. She is currently Assistant professor in the Information technology department at Madhav Institute of Technology & Science, Gwalior, India. Her area of research is Mobile Ad hoc networking, Wireless Network, and Image Processing.

