A Survey Paper on Techniques and Applications of Web Usage Mining

Subhi Jain
Dept. of Computer Science & Engineering, Graphic Era Hill University
Dehradun, India
2510shubhijain@gmail.com

Ruchira Rawat
Dept. of Computer Science & Engineering, Graphic Era University
Dehradun, India
ruchira.rawat@gmail.com

Bina Bhandari
Dept. of Computer Science and Engineering, Graphic Era Hill University
Dehradun, India
kotiyalbina@gmail.com

***Abstract*— The process of finding out valuable knowledge drawn out from web data is known as web mining. Identifying the various patterns and utilizing the vast knowledge extracted from those patterns is important from various perspectives such as business intelligence, e-learning, personalization etc. The web mining area which deals with extraction of patterns from user's weblogs is called as web usage mining, which is an implementation part of data mining. This paper focuses on the working of web usage mining, data sources for web usage mining and applications of web usage mining is explained in detail in this paper. Further, we explain the issues and current challenges in web usage mining.**

Keywords*— *Web Usage mining, pre-processing, pattern analysis, pattern discovery, personalization, privacy

I. INTRODUCTION

With the advent of various technologies, social networking sites, e-business site, e-learning sites etc. Every second new data is generated. The data growth rate is exponential. On a single tweet, for example, one can find millions of messages or retweets on Twitter. All this is possible because of the internet. Thus, The World Wide Web is a large source of varied data, a collection of billion documents, which either comes from the content in the web pages or from the various hyperlinks or structure of the websites i.e. web structure or from the log files depicting the web usage. The data or information is collected from various points or data sources. The Internet is also a form of network and in a network data is available at various nodes. Here, nodes can represent servers, client machines, intermediate devices popularly called as proxy servers or various databases that are stored on machines. Web mining is the part of data mining from which we derive meaningful and useful knowledge from text or contents present in web pages, hyperlinks and user usage logs [2]. For a clearer understanding, web mining has 3 parts: Web Structure Mining, Web Content Mining, and Web Usage Mining. The Web Content mining deals with the raw data available in the web pages; the source data mainly consists of the textual information, images, graphic, audio etc. present in the web documents. Since web content mining is basically related to the web content**,** the main application fields of web content are a content-based ranking of web pages and content-based categorization [3]. The main focal point of Web Structure mining is the structure of websites. The way in which a web page is arranged is outlined by the structure data. There are two types of structure information: Inter-page structure information and Intra-page structure information. The intra-page structure information is illustrated by a tree-like structure which is composed of HTML and XML tags. [3,9] whereas the hyperlinks that associate one page with another is inter-page structure information. With the information obtained from web structure mining, link-based categorization of web pages is done; this is an application of web structure mining. Web Usage Mining is the portion that focuses on extraction on knowledge from log files and helps in finding user behavior [13,15]; sources for these log files is servers, proxy servers or clients which can collect the typical information such as IP address, page references, access time etc. which are represented into standard formats and major application areas are web personalization, Business intelligence, E-commerce, E-learning etc.

The paper is detailed as Section I defines web mining and its types; Section II illustrates web usage mining, Section III discuss the process of Web Usage Mining and the various data sources. Section IV describes the different techniques used in the process of web usage mining. Section V discusses the application areas in detail. Section VI discusses the trending issues and challenges faced in this field. Section VII discusses the future trends in web usage mining.

II. WEB USAGE MINING

Data mining is implemented in many forms, web usage mining is one of them. Whenever a user interacts with the web pages, Usage mining does the task of finding the hidden important information about user behavior, its page surfing patterns and other valuable information which is used for various purposes. Data is obtained from various sources for web usage mining. Some of them are discussed below [2,4]:

A. Server Logs

These include the log files on the server-side for collecting information about the user such as IP address, access time, links visited etc[3]. Out of many weblogs, some acknowledged logs are as follows:

a. Common Log Format(CLF): The CLF is framed in order to monitor the chronological order of requests that occur on a website[4].The constituents of a CLF are shown through an example:

138.0.0.2 user-identifier frankin [11/jan/2000 :13:55:36 -0700] "GET/apache-pb.gif HTTP/1.0" 200 2326

- 138.0.0.2: The IP address of the remote host, also known as the client which has requested for a page from the server.
- user-identifier: This field represents the identity of the client.
- franklin: The document is requested from a particular user id, here, Frank is the user id.
- [11/jan/2000:13:55:36 -0700]: The time represents the time zone when the request has been received, the format is a date, month, year, time(H: M:S).
- "GET/apache-pb.gif HTTP/1.0": is the base URL or the request line of the client. GET is the method used by the client. Other methods are HEAD and POST. /apache-pb.gif is the resource requested, Name of the protocol is HTTP is the protocol used. 1.1 is the version of the HTTP protocol.
- 200: This is the HTTP status code representing successful response returned to the client.
- 2326: This is the size of the object in bytes returned to the client.

Extended Log format: Extended log format or ELF is more flexible and informative text file format. This is like CLF but has other extra details such browser details, the referring URL and the name of the host operating system. However, the data contained in these log files can be highly unreliable. The reasons behind unreliability of the data are IP address misinterpretation and Web caching.

Web caching: In this procedure, the main focus is on reducing the web latency. To reduce web latency, a copy of requested web pages is stored either in an intermediate server or in user's local browser. The requested page, if cached, is saved in the proxy server or local browser and therefore the server is oblivious to any new access to the page. Since the server is unaware of the access, the access is not recorded to the log files.

IP Misinterpretation: This is the second source of unreliability which occurs because of two reasons:

- Many users use the same computer, hence the IP is same.
- Allocation of the same IP to all the users because of the use of proxy servers.

b. Explicit User Input: This is another variant of data which is collected externally when the user accesses a website, often collected from user registration forms. This data is often incomplete and inaccurate.

c. Cookies: Cookie is a string or a small piece of data sent during the browsing of the web by the user, from a website or server's end and is stored on the user's computer (client side) by the web browser. The small size of cookies gives an edge to it, as compared to other data sources.

d. Click streams: A click stream is a sequential registering of what a user clicks on while surfing the internet website or when it uses particular web applications. A user session can be considered as the number of clicks recorded when a single user surfs the web.

B. Client Data

This is the second data source which is collected from the host that accesses the websites or requests for the web pages. Remote agents are implemented in Java or javascript to access the data. They collect information like user's navigational history. It is more reliable than server data as it has no problem with web caching and IP misinterpretation.

C. Intermediate Data

When the user's request is granted through proxy servers or intermediate devices, intermediate data is generated. Therefore, proxy servers and packet sniffers also act as a source of data.

a. Proxy Server: A proxy server is an application or program that acts as an agent to grant the requests made by the client while it is looking for resources from other servers [20]. An access log is employed in a proxy server to store the web page requests made to the server and the responses received from the server[4]. An access log is defined as a file, a list or a group of files which has a list of each and every file or a page that the user has requested from a server or a particular website. The access log files generally include associated files with the web pages that get transmitted, HTML files, some graphics etc[19].However, issues of web caching and IP misinterpretation exist here also.

b. Packet sniffers: A program or a hardware that checks the flow of data in the network is called a packet sniffer [21]. The TCP/IP packets that are directed to the web servers act as the source of data. Generally, log files do not contain any time stamp information such as request or response time.This information is present in packet sniffers. Data loss is a problem in packet sniffers as data is not logged anywhere.

III. WEB USAGE MINING PROCESS

The process of web usage mining comprises of the following steps [2]:

- Data cleaning
- Data pre-processing
- Pattern Analysis
- Pattern discovery

Fig.1 is the diagrammatic representation of web usage mining phases. Further, every step is discussed in detail as follows:

A. Data Collection

The first step is a collection of data from varied sources. Apart from minor sources, the major sources are: (a) Web servers, (b) Client machines, and (c) proxy servers [4].

a. The server side: The server-side has a substantial amount of information, which it represents in a standard format like CLF or ELF, stored in server log files or the database log files [3, 4].

Issues: The major issue is user session identification i.e. identifying the click stream and the path followed by the user during visiting that website [3]. Even if we use cookies to remove this issue, the tracing of back button navigation is not performed at the server level, which makes it difficult to find the exact path followed by the user during its web browsing session [3,4].

b. *The proxy side*: These are intermediate servers which improve navigation speed through caching and they collect data from huge servers by making groups of users. The collections of log data from proxy servers are similar to server-side data collection in many respects.
Issues: since there may be frequent caching between users and proxy servers it is difficult to reconstruct the session as the path of all the users' navigation cannot be identified [3].

c. *The client side*: The client-side data is the log files present on the user side. Here, we can track the usage data on the client side. JavaScript, Java applets and modified browsers are used to trace the data on the client machine. The problems of user identification and session identification are dodged [3].

Issues: Every machine may not fully support Javascript or Java applets to track the data.

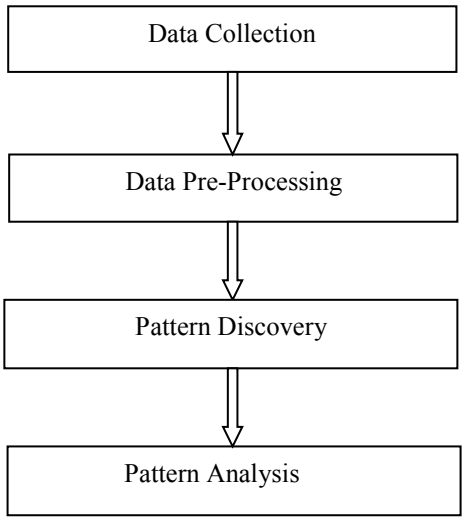

Fig 1: Phases of Web Usage Mining phases

B. Data Preprocessing

After collection of huge amount of data, it is necessary the data is consistent, integrated and relevant. Thus, Data needs to be ready for pattern discovery and analysis which is known as pre-processing of data [5]. It is a time consuming and intricate process. It includes four different tasks: Data Cleaning, session identification and user identification with the rebuilding of user's session, Recovering the information pertaining to the content of the page and structure and the data formatting [3,4, 5, 8].

a. *Data Cleaning*: The data collected on the internet has information which is not used as a request for graphical page content (images in .jpg and .gif format), ads etc [3]. Elimination of useless data from weblogs is the primary job of this step. A weblog is a website that has series of entries (logs) of the pages or sites or links surfed by the user. This is performed by spiders and robots [3].It is easier to drop graphical requests but the navigation patterns of robots and spiders are explicitly identified and removed. By keeping a track of access to robots.txt file these patterns can be removed effectively [3,4].
Another approach used for robots that use false user agent in HTTP session from actual approach is heuristic based (i.e. either on a classification of models or presumptions). The classifier is trained with the help of well-known robot's navigational path and the further categorization is done using the acquired [3].

b. *User and Session identification and reconstruction of user session*: In this step, we find the user's session from the weblog data obtained and as the next step within the identified session, rebuilding of user's navigation route is performed. For session identification, we use cookies, URL rewriting etc. The various user and session identification methods are discussed below.

***User Identification*:** It is the process of identifying the users' on the web i.e. which user accesses the web pages or requests for the websites or web pages. There are various approaches to automating user identification [1,22]:

- A unique user id can be given to each user in a log file; this is one of the methods of identifying the user. However, if the user uses same machines or proxy servers it is difficult to identify the user.
- Cookies are used for finding the user but are subject to deletion of disability.
- Another method is of the various special services present on the internet like in or fingered services for user recognition
- Even if the IP of the user is same, by analyzing the various weblogs the difference in browser types, operating systems, and their versions are identified.
- The topology of the site can be merged with access log entries to identify if the user is new or not. This is distinguished by the fact that if access with an identical IP is made to an identical page without a direct hyperlink between them, then the user is a new user.
- Instead of cookies, the web server includes a unique ID in the URL and the user creates a bookmark for one of the delivered pages in order to gain access to the website. Thus, this method is not fully automatic as user needs to bookmark the page first.

Session Identification: The information of the navigation pattern of the user is put into code by session identification. User sessions can be found out with the help of various proposed method which is either based on time or based on the content [1, 22].

Time-based methods: Most of the models use these methods and have several ways as discussed below [1, 3, 22]:

- A default 30 mins time period is assigned. All the web pages accessed in this time period are considered as a single user session.
- In the next approach, the client sends it's time to the machine on server side every time a fresh page is contained. Java agents help in this process. This approach is influenced by the availability of JavaScript, the variant of browser and the traffic in the network.

Content-based methods: In content-based methods, the sessions are computed based on the transactions to offer valuable things. Here a transaction can be said as a subset

of pages that befall during a session of a user. The various methods are:

- The web pages are mainly divided into three parts: content pages, navigational pages and user and hyper pages. Hyperlinks to one page from another are present in navigation files. The real information or data is present in content pages. The pages that contain both navigational and content information are user and hyperweb pages. This categorization depends on user's need and is very stern.
- A page can be categorized as a content page or a navigational page depending on the time spent on that particular web page by the user. If the browsing time on the web page is greater than a certain period, the page is called content page else it can be called a navigational page.
- The content page is the last page visited in a browsing session.
- The set of pages visited during a browsing session starting from the first page visited the first backward reference is called a transaction. The next forward reference is the start of a new session. Web caching is an issue with this method as backward references are not recorded.

c. *Content and structure retrieving:* URLs are generally poor sources for content retrieval. Therefore to enrich web log data a technique of content-based information was employed. The weblogs make use of the classified information generated from the web pages. Web structure mining is employed to develop an adequate classification technique [11].

d. *Data Formatting*: prior to applying data mining techniques the data must be properly formatted. We can store the extracted data into the relational database, a tree-like structure or cubical structure can also be used [3].

C. Pattern Discovery

Pattern discovery defines various methods and algorithms that are used to extract valuable patterns by analyzing the pre-treated information [10]. The various methods include pattern recognition, data mining, statistical analysis and machine learning. Statistical Analysis, sequential pattern analysis, clustering algorithms, association rules algorithms, classification techniques and dependency modeling are some techniques used for pattern discovery. These will be discussed in detail in section IV.

D. Pattern Analysis

After the discovery of patterns from pre-processed raw data it is necessary to find interesting patterns and filter out unwanted information, this task is performed by pattern analysis. There are various methods used for pattern analysis as discussed below:

- Various knowledge retrieval mechanisms that help in finding out valuable information with help of queries in structured form is used.
- OLAP operations can be used for patterns analysis.

Visualization techniques can also be used to highlight patterns or graphs can be constructed to demonstrate the analyzed patterns.

IV. TECHNIQUES

There are various techniques for the analysis of web usage data and discovering knowledge from it. The major techniques are as follows:

A. Statistical Analysis

This is the most common technique used in which a session file is analyzed and statistical operations like mean, median, frequency etc. Can be applied on navigational paths, browsing routes, visited pages etc.

B. Association Rules

Association means a relation that exists among entities. The rules of association were first instituted by Rakesh Agrawal in 1993.This approach is the most commonly used in web usage mining. In this approach, our goal is to find the relationship amongst a various set of items. It can be represented as (item1)=>(itemm2), where item1 and item2 are set of items in a set of the transaction. According to association rules, the above notation states that if a purchase or a transaction contains item1 it is highly likely that it will contain item2[3,15]. For e.g.:

"Apple, Bread=>Butter"

This states that if a person has brought apple and bread in the same session the same person will buy butter also. The association rules have the basic fundamental of *support* and *confidence [6]*.

Support: It tells how popular a particular item set is, measured by the proportion in which it seems. These items in the item set which are popular have sales beyond a particular support threshold called minimum support.

Confidence: It tells how likely we will buy Y item if x item is bought. This also has a minimum confidence which is the threshold value.

For e.g.: Consider a grocery store transaction database:
purchase (T;"Apple")purchase(T;"beer")

Here, T is the customer. With the above transaction, support and confidence values are also associated. Let say confidence is 80% and support is 5%, this means that if Apple is bought by the customer there are 80% chances that beer will be bought. The support value means, that among all the transactions, 5% of the transactions in the database show that beer and apple are purchased together as a single dimensional unit.

The main goal in association rule mining algorithms is to generate candidate item-sets and frequent item-sets. The candidate item-sets are the item-sets that have a hope to be large or frequent item-sets. Frequent item-sets are the item-sets with minimum support threshold value.

The algorithms are:

- Apriori algorithm
- Reverse Apriori Algorithm
- Rapid Association Rule mining
- FP-growth Algorithm
- Eclat Algorithm

C. Clustering

In this technique, the group of items having similar properties is clustered or grouped together. Usage clusters and page clusters are two kinds of clusters. Making clusters of users is to set up the group of users which show the similarity in browsing patterns. In contrast, Page clusters will find a group of web pages on similar content [23]. Clustering is categorized into three types: partitioning methods, hierarchal methods and model-based methods.

a. *Partitioning methods*: K means clusters or groups of data are formed in portioning method. Algorithms are applied depending on the purpose and need of the user.

User session clustering algorithms:

- K-means clustering algorithm
- K-means with genetic algorithm
- Ant-based
- Self-Organising maps
- Graph partitioning

Index page synthesis algorithms:

- Page gather

b. *Hierarchical methods:* Hierarchal methods create a hierarchal structure of clusters by breaking the data into subgroups. The algorithm for clustering user sessions is BIRCH [1]. This algorithm handles 'noise' effectively.

c. *Model-based methods:* in this method, best fit among the datasets and mathematical sets is found out. User session creation algorithms are:

- Auto-class
- Self-organising
- COBWEB
- ITERATE

D. Classification

In classification, we map the various data item sets into several pre-defined classes [14]. The characteristics of a class are defined by the choice of features and the features extracted. It is done by supervised inductive learning algorithms. For e.g. classification algorithms are applied to log files to generate useful rules like 50% of the people when browsing the internet open facebook and Instagram simultaneously and are aged between 18-30 and browse during 12:00-3:00 pm.Some classification algorithms are[1, 18, 19, 22, 23, 24, 25]:

a. Representing user interest and extraction rules algorithms:
   - HCV
   - CDL4

b. Session classification based on the concept:
   - Rough Set Theory

c. Interesting page prediction algorithms :
   - C4.5
   - Naive Bayesian

E. Sequential patterns

In this technique, inter-session patterns are found out, such as which set of items follow another set of items in a particular time session [9]. In this technique, future visit patterns are predicted by web marketers [1, 9]. Methods to draw out sequential patterns are listed as deterministic methods and stochastic methods.

a. Deterministic methods: These methods record the navigational behavior of the user. There are several algorithms for this approach [1,22]:
   - Spiliopoulou: Used in sequence rules extraction.
   - CAPRI: used in clustering of navigational patterns.

b. Stochastic methods: Subsequent behavior in this method is predicted by the order in which the web pages are visited. The algorithms used are:
   - Borges: It makes use of user session to extract navigation behavior of the user and generate the patterns.
   - Markov model: This algorithm is used for prediction of next-link by the user or can be said as the expected browsing by the user.

F. Dependency Modeling

In this technique, a model showing important dependencies is developed. [9].for e.g., the various stages a user undergoes when he starts learning a course online till the finish of the course are included in this model.

V. APPLICATIONS

Valuable patterns are drawn out from the outcomes produced from the web usage mining process. The applications of results are [7]:

A. Personalization of web content

Personalizing the experience of the user is important from the business perspective of many web-based applications. Personalization means showing those results to user, which the user wants to see based on his navigation patterns, search patterns and other patterns discovered from web log data[1]. For example, through users browsing pattern we can identify the type of web pages and websites the user visits and customize the web for him using this analysis.

B. Pre-fetching and caching

The server response time, access time is cut using the caching and pre-fetching technique by means of using the outcomes extracted by analysis of weblogs.

C. Support for the design

To make the websites efficient and flexible to use by all the users, weblog extracted data patterns helps in design issue of the website.

D. To improve customer satisfaction

With personalized experience the user gets a better access to the website and an improved experience on its features, hence improving customer satisfaction.

E. Business Intelligence

Drawing out of knowledge from information is a primary task performed by web usage mining and determines customer behavior. It helps in determining effective marketing strategies that help in increasing the sales [10,28].Business intelligence is all about helping people make good decisions and maintaining competitiveness in the marketplace.

F. Enhanced E-learning

The course's activities can be traced using the web usage data and suggestions for system improvement are placed using this information [27].

VI. CURRENT ISSUES AND CHALLENGES

The major issues in web usage mining are discussed below:

- Privacy is considered as one of the most major issues in web usage mining [7]. Since data is collected from vast and varied data sources like cookies, weblogs, URL's etc. it is difficult to support the privacy of data. The solution to privacy problem is P3P i.e. privacy preference standard which proposes various privacy standard format [12,26].
- Another problem is dealing with large and vast volumes of data which is exponentially growing with time.
- The lack of information on the comparison of various tools which makes evaluation criteria difficult.

VII. FUTURE TRENDS

Web usage mining has various issues which give open research areas in this field which help in developing future trends in this domain.

There are two prominent issues in this area as discussed above. Firstly, privacy is a big challenge. Secondly, integration of semantics within websites, this is also an open research area, which is also a research area [3]. There are several approaches followed to develop trends like modeling. The future trends will mostly revolve around privacy and semantics. Thus, we have various open research area in web usage mining which gives rise to future trends.

VIII. CONCLUSION

Due to vast extensions of the network, various websites and applications have emerged that keep the extensive amount of user information. Web usage mining helps in discovering the patterns from user's weblogs and through that extracted information, knowledge is drawn out and used in various fields like e-commerce, online business sites etc[6]. Studying and analyzing this information helps the designers in catering to user's specific needs i.e. personalizing user experience and efficiently organizing the websites. Also, customer behavior patterns help in establishing business rules and based on these the organization identifies the future needs the customer may develop and their current likings [3].In this paper, we presented a survey on web usage mining, the processes involved, the techniques used. Further, we have discussed various issues and challenges and the future trends which are mainly related to privacy and dealing with large volumes of data.

REFERENCES

[1] Dimitrios Pierrakos, G E Orgios Paliouras, Christos Papatheodorou, and Constantine D. Spyropoulos, "Web Usage Mining as a Tool for Personalization: A Survey", User Modeling and User-Adapted Interaction 13: 311^372,2003.

[2] Priyanka Bharti, Dr.Sona Malhotra, "A Review Paper on Web Usage Mining and future request prediction", Special Issue (ICFTEM-2014) May 2014 pp. 33-37 (ISSN 0973-4414) 33.

[3] Federico Michele Facca, Pier Luca Lanzi, " Mining interesting knowledge from weblogs: a survey", Data & Knowledge Engineering 53 (2005) 225–241.

[4] 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.

[5] Mitali Srivastava, Rakhi Garg, P. K. Mishra, "Preprocessing Techniques in Web Usage Mining: A survey", International Journal of Computer Applications (0975 – 8887) Volume 97– No.18, July 2014.

[6] Dr. G. K. Gupta, "Introduction to Data Mining with Case Studies" PHI Publication, 2005.

[7] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1, No. 2, Pp. 12-23, 2000.

[8] Tanasa, D.; Trousse, B.; "Data preprocessing for WUM", IEEE Potentials, Vol. 23, No. 3, Pp. 22 – 25, 2004.

[9] P.Nithya, Dr. P.Sumathi," A Survey on Web Usage Mining: Theory and Applications", Int.J.Computer Technology & Applications, Vol 3 (4), 1625-1629.

[10] Cooley, R.; Mobasher, B.; Srivastava, J.; "Web mining: information and pattern discovery on the World Wide Web", Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence, Pp. 558 – 567, 1997.

[11] R. Cooley, B. Mobasher, J. Srivastava, "Data preparation for mining world wide web browsing pattern", Knowledge and Information Systems 1 (1) (1999) 5–32.

[12] Platform for Privacy Preferences (P3P) Project, http://www.w3.org/TR/P3P/ (2003).

[13] Bina Bhandari, Bhaskar Pant, R H Goudar," ARAA: A Fast Advanced Reverse Apriori Algorithm for Mining Association Rules in Web Data", International Journal of Engineering and Technology (IJET).

[14] Bina Kotiyal, Ankit Kumar, Bhaskar Pant and R. H. Goudar," Classification Technique for improving user Access on Web Log Data", Advances in Intelligent Systems and Computing 243, DOI: 10.1007/978-81-322-1665-0_111, Springer India 2014.

[15] Bina Kotiyal, Ankit Kumar, Bhaskar Pant, R H Goudar, Shivali Chauhan, Sonam Junee," User Behavior Analysis in Web Log through Comparative study Of Eclat and Apriori".

[16] "Mining Association Rules between Sets of Items in Large Databases" Rakesh Agrawal, Tomasz Imielinski_, Arun Swami; Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA, May 1993.

[17] "Association Rule Mining: A Survey", Qiankun Zhao and Sourav S. Bhowmick Nanyang Technological University, Singapore, Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116, 2003.

[18] D. Dixit *et al*., "Mining access patterns using classification", *International Journal of Engineering Science and Technology*, vol. 2, 2010.

[19] Searchsecurity.techtarget.com/definition/access-log

[20] https://www.techopedia.com/definition/4113/sniffer

[21] K. S. Reddy, G. P. S. Varma, and S. S. S. Reddy, "Understanding the scope of web usage mining & applications of web data usage patterns", in *Proc. International Conference on Computing, Communication, and Applications*, 2012, pp. 1-5.

[22] M. Valera, "The descriptive study of preprocessing from web usage mining", *Indian Streams Research Journal*, vol. 2, pp. 1-6, 2012.

[23] Chitraa and A. S. Thanamani, "An enhanced clustering technique for web usage mining," *International Journal of Engineering Research &Technology (IJERT)*, vol. 1, p. 5, 2012.

[24] S. P. Nina, M. Rahman, K. I. Bhuiyan, and K. E. U. Ahmed, "Pattern discovery of web usage mining," in *Proc. International Conference on Computer Technology and Development*, 2009, pp. 499-503.

[25] S. K. Pani, L. Panigrahy, V. H. Sankar, B. K. Ratha, A. K. Mandal, and S. K. Padhi, "Web usage mining: A survey on pattern extraction from weblogs," *International Journal of Instrumentation, Control & Automation (IJICA)*, vol. 1, 2011.

[26] Q. Zhang and R. S. Segall, "Web mining: A survey of current research, techniques, and software," *International Journal of Information Technology & Decision Making*, vol. 7, pp. 683-720, 2008.

[27] O. R. Zaïane, "Web usage mining for a better web-based learning", presented at the Conference on Advanced Technology for Education, 2001.

[28] Q. Zhang and R. S. Segall, "Web mining: A survey of current research, techniques, and software," *International Journal of Information Technology & Decision Making*, vol. 7, pp. 683-720, 2008.