Web Usage Mining Algorithms: A Survey

Sowmya H.K.[*] **, Dr. R.J. Anandhi** [**]

[*] Department of Computer Science and Engineering, The Oxford College of Engineering, Bangalore, India
hk.sowmyakiran@gmail.com

[**]Department of Information Science and Engineering, New Horizon College of Engineering, Bangalore, India
rjanandhi@hotmail.com

***Abstract*—**The rapid progress of WWW has created gigantic amount of web data. Web usage mining can be applied to determine valuable information from the user communication with the WWW. The main objective of mining of web usage is to interpret the response of web site customers via the approach of data mining. Insight gained from web usage mining perhaps employed to boost designing of website, grant personification service and clear the way for more productive browsing. To attain this, it is required to secure pattern of user access in the form of log files. Web Usage mining is a way of determining communication of user with distinct web application. It is a process, which mainly consists of three inter-dependent phases like preprocessing of data, discovery of pattern and analysis of pattern. This survey paper objective is to study various techniques and algorithms for data preprocessing and data discovery stages. The authors have identified the gaps that are present in various web usage mining algorithms and proposed techniques to handle few of them. They also studied and analyzed frequent itemset mining algorithm like Apriori, FP-growth and SSIFM (Single-scan Frequent Itemset Mining Algorithm) on web usage data.

***Keywords*—** Data mining; web usage mining; preprocessing; pattern discovery; pattern analysis

I. Introduction

Mining of data is a method of extracting knowledge from massive volume of input. Repository of data perhaps in the form of database, data warehouse or data marts. Process of data mining involves three major stages specifically data pre-processing, pattern extraction and pattern analysis to discover useful patterns.

Web mining is a branch of data mining, helpful in discovering hidden pattern form large volume of web data repository. World Wide Web is an information system which holds huge data in the form of links, images, audio, video and graphics files. Web mining is somewhat distinct from data mining because it majorly pertain to unstructured or semi structured data, while data mining concerned with structured data.

Web mining conceivably on the whole subdivided to form three categories: The headmost subgroup is web content mining. It tries to obtain utility facts or comprehension from the content of web page. Web page may consist of text, audio, video, image, metadata and hyperlinks etc. Investigation in web content mining comprises acquiring facts from the web, classification of document and grouping, and drawing out information from web pages. The second subgroup is web structure mining. It tries to uncover knowledge from the hyperlink structure. It is categorized in to two kinds such as intra page structure and inter page structure. In Intra page structure, same page holds the link. It does not support opening of different page. Inter-page structure assist linking of one page to some other page. The third subgroup of web mining is called as web usage mining. It relates to the exploration of browsing pattern of user from web log. It gives prominence for diverse techniques of data mining to perceive and interpret identified patterns.

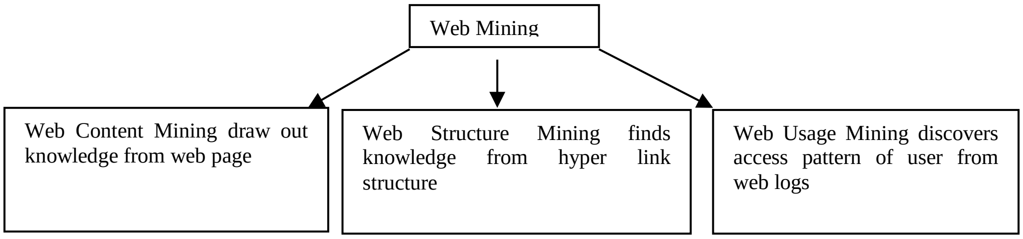
Web Mining
Web Content Mining draw out knowledge from web page
Web Structure Mining finds knowledge from hyper link structure
Web Usage Mining discovers access pattern of user from web logs

Fig. 1: Categories of Web Mining

Web Usage Mining

Web usage mining pertains to the preprogrammed exploration and investigation of patterns in web log. Discovered usage pattern is helpful to perceive and more effectively support the requirements of Web-based applications. The main purpose is to acquire, design, and analyze the user access patterns and their profiles, when they communicate with website. In mining of web usage, several approaches are employed to extract data produced by proxy server, web server and cookies to arrive at navigation pattern and browsing nature of the user.

Fig. 2 shows the three distinct phases of Web Usage Mining. Preprocessing of data is the first phase under this category. The web log file consists of huge amount of incomplete and unnecessary information, so straight away employing of such primitive log to perform web usage mining is not possible. Quality of web log file can be improved by applying preprocessing technique. Different techniques are applied in preprocessing phase that is cleaning of data, identification of user, sessionization, data integration, data transformation.

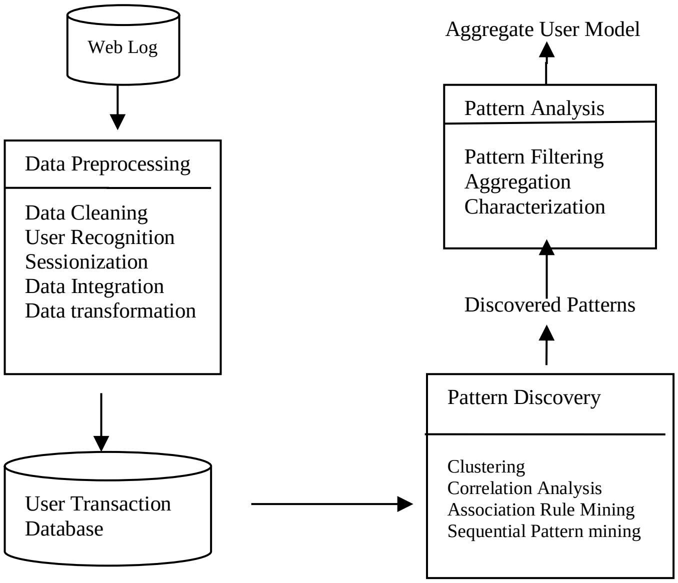

Fig. 2: Web Usage Mining Architecture

A. *Data Cleaning*

It is a mechanism to recognize and eliminate rows from web log which are not necessary and conformant. Web log consists of several attribute fields, from that only necessary fields are selected and others are removed. First step in data cleaning is to remove records with file extension JPEG, GIF, CSS and so on, as the web pages are not of user interest, and instead it is implanted objects in the web page. Second step is to discard the entries which are marked with error status code, in case of user request for the web page which is not available on web server. Third step is to remove the web log entries gathered from the crawlers or machines as they do not reflect the way in which actual user move across the website. Several search engines announce them as an agent and thus can be recognized without any difficulty by string matching.

B. User Identification

It is a process to find out the user one who contacted web site and which pages are secured. If users have registered their details, then identification of such users is easy. There are several users who did not sign up their information and access web sites via, agent. Moreover, the same person can use different systems, and different users can use the same system as well. Additionally, proxy server conceals suitable facts about individual users as different computers take part on the internet work with the same IP address with the help of proxy server. All the above complications make the effort of user identification greater difficulty and complexity. Cookies can be used to correctly track users' behaviors. But taking privacy in to account, many users do not use cookies. Hence it is required to determine other approaches to provide solution to this problem

C. Session Identification

Sessionization, is the assignment to recognize the sessions from the raw data. But the challenge here is that, the log file of server do not hold entire essential facts. Modify data in to a suitable form, so that they can be treated as data for the algorithms. Data mining techniques will be applied to the appropriate data to achieve predetermined goal.

The second stage of web usage mining is Pattern Discovery. After identifying user sessions, apply various mechanisms and process originated from various fields namely data mining, machine learning, statistics and pattern recognition. Based on the requirement of the investigator, different kinds of pattern access, such as finding association rules, path analysis, discovery of sequential patterns, and clustering and classification can be performed. Identifying of desired patterns and to extract understandable knowledge from them is a challenging task.

Analysis of pattern is the last phase in the web usage mining. Main objective of this phase intend to remove the guidelines or arrangements which are not relevant. And also retrieve the relative rules or patterns from the outcome of pattern discovery phase. Structured Query Language (SQL), Online Analytical Processing (OLAP) are the tools used to perform pattern analysis.

D. Data Sources

There are three important data sources for web usage mining which includes data repositories such as 1. Server log 2. Client log and 3. Proxy log.

Server log provides details corresponding to page asked by the user. Details of the current requests are entered at the end of the file. This log file holds information related to the request made by the user such as, IP address, date and time, URL of requested page, HTTP code, bytes served, user agent and referrer. All this input conceivably aggregated in to one file or segregated in to various log files, such as error, access and referrer log. Server log file is projected as not complete file since it do not document the cashed pages. These are mainly user admin purposes such as, constructive web site administration, established resources management.

The browser window of the client holds client log file within it. Recording of details to the log file are made by the Web server. These files are used to represent user behavior as they are most genuine and precise. But it is complex exercise to alter the browser for each user and demands users' concentration and cooperation.

Proxy log file hold the HTTP requisition from various users to various Web servers. This log file is a source of data to uncover the pattern of use of unidentified users who share a same proxy server. As Proxy server log files are more susceptible and composite, it is more difficult to deliver true picture user behavior.

II. Literature Survey

The best approach for preprocessing of web usage mining is proposed by K. Sudheer Reddy [7]. Authors have proposed algorithms for user and session identification. This method not only decreases the volume of the log file further improves the quality of data. Mitali Srivastava et.al [3] made an observation of extraction of data and cleaning of data in web usage mining. They suggested data extraction and data cleaning algorithm, which works on web log file. Result of the experiment in this paper showed that raw log data reduced to 80% after cleaning process.

Neha Goel et.al [4], designed a tool for preprocessing web logs. It is a complete tool for reducing data which are not relevant and appropriate. And also modify it into a suitable form so that it can be applicable for analysis. The proposed methodology also figures out some enhanced attributes in statistics form namely the amount of time that has passed in getting the outcome, hit count by corresponding IP. K. Sudheer Reddy et.al [10] developed algorithms for retrieving log file from web server and combining log file. In the same paper, authors also focused on the preprocessing approaches realized on Web Sift tool.

B. Uma Maheswari et.al. [5], suggests Greedy clustering algorithm using belief function for user profile creation. This algorithm uses belief function similarity measure to assist clustering task that has capability to apprehend the unreliability between web user's navigation attainments. This study presented a method based on referrer information for making error free transactions in data preprocessing. Outcome of review is provided to the group of routines which makes use of Dempster's rule.

P. Nithya et.al [9] presented preprocessing approach for mining web log to eliminate noise and robots in the web. The output of preprocessing is an error free input for the later stages of data mining. For measuring the proposed preprocessing technique anonymous web data set is used, and it discloses the number of records. P. Dhanalakshmi et.al. [1], developed algorithms for static log cleaning static user and session identification.

Priyanka S. Panchal et. al. [8], proposed composite method for predicting users acquire pattern rest on Markov model. This technique performs grouping of user exploration build on their affinity measure along with concept of Markov model. Here the fundamentals of apriori algorithm can be set for guessing web link. Web link prevision is a way to predict the user visit to web page based on their past visiting record.

Yakhchi S et.al. [12], proposed improved ARMICA which deals with numerous attributes such as the count of accomplished rules, database scans, and the aspect of produced rules. Authors compared the novel approach with the existing algorithms and outcome of the experiment showcased, ARMICA-Improved is quicker, and produces reduced count of rules with more aspects. It has taken reduced count of database search to produce more exact results.

Enhanced hybrid system proposed by Janisa Colaco et.al. [14], for predicting user navigation pattern. These novel algorithms are used for locating chronic pattern and grouping the identified patterns. This hybrid algorithm shows enhanced results compared to the existing system based on Apriori. Smaller amount of time is taken by this technique, which resulted in improvement on computational and clustering efficiency. Mining Least Association Rule (MILAR) algorithm [20] is applied on student exam dataset to obtain the secondary association rule. Inference shows that the count of implied association rules are downsized as there is increase in CRS (Critical Relative Support) value. Further it reduced the number of rules that are apathetic.

Abdelghani Guerbas et.al [22] applied DBSCAN and OPTICS clustering algorithm for productive mining of web log and prevision of navigational pattern accordingly. Authors also recommended an outline for the best online navigational behavior prediction. It assists in decreasing the reaction time of server by accumulating pages that are most probably appealed by a user.

III. GAPS IDENTIFIED IN EXISTING DATA PREPROCESSING ALGORITHMS

An algorithm for static log cleaning, static user and session identification by P. Dhanalakshmi et.al. [1], has greater attainment in terms of error rate and F-measure only on restricted data volume and fields. Session identification algorithm proposed by P. Sukumar et.al [2], has not explored accuracy metric for the algorithm to identify user and session.

Precision benchmark for user and session identification algorithm is not tackled in preprocessing method proposed by K. Sudheer Reddy et.al. [7], for web usage mining. Effectiveness of the method proposed in this paper is not measured. Algorithms recommended for cleaning data, data fusion and data extraction by Mitali Srivastava et.al [3], are not efficient as they are inappropriate for huge data set. Time taken by algorithms is increasing for large data set. So the proposed algorithm does not have the ability to scale up with growing data set.

The tool that was designed for preprocessing web logs presented by Neha Goel et.al. [4], is restricted to a few records and to a specific website. So it is required to test the algorithms proposed by authors in this paper with different website for measuring its performance. K. Sudheer Reddy et.al [10], suggested algorithms for extracting and joining log file from web server. These algorithms require web log files of longer duration and more rules to improve server performance.

Greedy clustering algorithm using belief function, suggested by B. Uma Maheswari et.al [5], is experimented on web log of smaller size. Only few number of user identified and session formed by this method. As this algorithm cannot scale up to work on huge web log, its usage is limited to small data set. Algorithm proposed by Nisarg Pathak et.al.[6], examines only unchanging log files. It needs to enhance in identifying navigation pattern of user for unsteady website.

New clustering technique proposed by Priyanka S. Panchal et. al [8], computes fixed number of clusters which affects the accuracy of web page prediction. Accuracy and efficiency of the preprocessing algorithms for discarding provincial and widespread noise and web vehicles, is not measured in the work stated by P. Nithya et. al. [9]. In this paper, author has not compared time taken for user interested pattern prediction with other widely used prediction techniques.

IV. GAPS IDENTIFIED IN EXISTING DATA DISCOVERY ALGORITHMS

Efficiency of the algorithm proposed by D.Kerana et.al. [11], is similar to Apriori algorithm. Imperialist Competitive Algorithm presented by Shahpar Yakhchi et. al.[12], requires a predefined parameter and a proper mechanism to set this value. Frequent item set mining using pattern growth approach developed by Rashmi et.al [13], is not tested for large volume of data.

Janisa Colaco et. al [14], developed hybrid algorithm for predicting pattern of user behaviour is tested only for EPA data set. The experimental outcome of the newly introduced system is not distinguished with various alternative current systems. Suggested algorithms by the author perchance evaluated with the help of other set of data. Additionally, the suggested system can be exercised to various operation fields to evaluate its performance and applicability.

Comparison of Dclat algorithm with other algorithms suggested by Vijayarani et. al [15], is not measured the quality of association rule. Novel fitness functions defined using multi-objective GA-PSO [16], need to be scrutinized for improving the kindness of the rules. Manju et.al. presented Ant Colony Optimization approach is employed to produce association rule in only one step [18]. In this paper association rule’s quality is measured only with respect to high confidence. New improved FP-tree algorithm developed by Ashika et.al. [19], tested only for fixed number of transactions of size 1000.

Algorithm proposed by Zailani Abdullah et. al.[20], not tested with other benchmark data set. Scalability of the association rule mining algorithm is not addressed in the Association rule mining using Apriori algorithm [21]. Abdelghani Guerbas et.al [22], applied DBSCAN and OPTICS clustering algorithm for persuasive mining of web log and anticipation of online navigational pattern. It uses two-step process for mining navigational patterns, OPTICS and then DBSCAN instead of combined density-based clustering algorithm.

V. PROPOSED METHOD

Collect user log data from web log of the servers, perform data preprocessing for data reduction, user identification, session identification. Design an efficient pattern discovery technique for grouping the web users based on their behavior (click streams). Develop an optimized clustering technique to group the users with similar browsing preference and web pages that appear to be theoretically associated with respect to the users’ perception. Develop an organized and qualified mining method based on association rule to find the co-occurrence relationship among the identified patterns to extract information. Efficiency factors to be considered are time factor and number of database scans.

VI. RESULTS AND DISCUSSSION

The present work carried out in Java Eclipse platform, for the study and analysis of Apriori, FP-Growth and SSFIM frequent itemset mining algorithm to structure highest-n frequent item set. Experiment is conducted on the web server log file of msnbc.com crunched from the website, "http://kdd.ics.uci.edu/databases/msnbc/msnbc.html". This website contains pre-processed web log data that reports the page visited by users on September 28, 1999.

Numerous experiments have been performed applying two types of data sets to examine the algorithm's performance for frequent item set mining such as Apriori, FP-Growth and SSFIM. The first dataset is a collection items, with average size of 4-16 having number of transactions 693. The second instance is a collection of medium-sized database instances, with 3196 transactions. All three algorithms in the experiments have been realized in java Eclipse Platform and experiments run on a desktop machine equipped with Intel I7 processor and 4GB memory. The following table present execution time of various frequent item set mining algorithm for MSNBC data set by varying the values of minimum support.

Table 1. Execution Time of Frequent Itemset Mining on MSNBC Dataset

MSNBC Dataset			
Min Sup	Execution Time in ms		
	FP Growth	Apriori	SSFIM
5	23	41	39
10	21	37	33
15	19	32	32
20	19	31	31

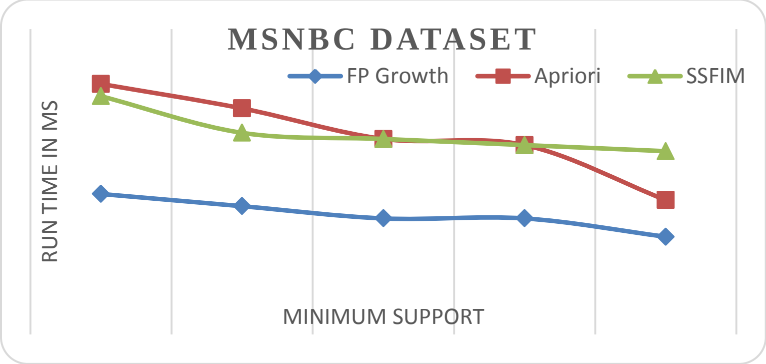

Fig. 3: Execution Time of Frequent Itemset Mining on MSNBC Dataset

Table 1 shows that, for dataset with small instances, FP-growth, Apriori algorithm surpasses SS-FIM. However, for dataset with large number of instances, SS-FIM clearly surpasses FP-Growth and Apriori. Fig. 3 shows the runtime performance of the FP-growth, Apriori and SSFIM using the msnbc datasets described above.

Table 2 approves that SS-FIM algorithm is finer than FP-Growth and Apriori when we handle sparse and large dataset. It presents execution time of various frequent item set mining algorithm for chess data set by varying the values of minimum support.

Table 2. Execution Time of Frequent Itemset Mining on Chess Dataset

Chess Dataset			
	Execution Time in ms		
Min Sup	FP Growth	Apriori	SSFIM
60	926	123419	99
70	310	19829	91
80	178	3872	83
90	138	485	77

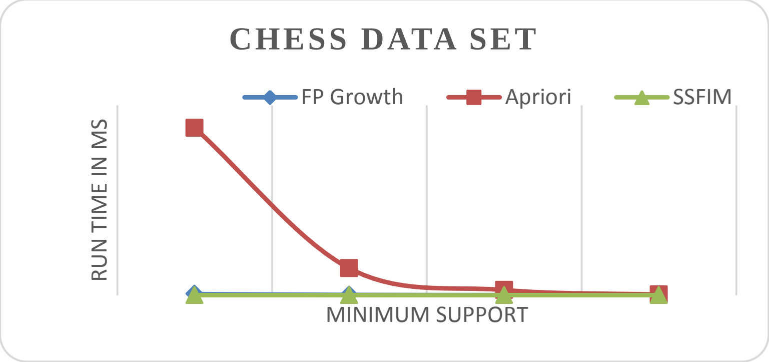

Fig. 4: Execution Time of Frequent Itemset Mining on Chess Dataset

Fig. 4 shows the runtime performance of the FP-growth, Apriori and SSFIM using the chess datasets described above. It shows that SS-FIM is not sensible to any variation of minimum support threshold value. The number of item set generated by this algorithm is fixed. Alternatively, Apriori algorithm performs multiple scanning of the dataset. It produces enormous candidate itemsets when minimum support threshold value is fixed with low value which declines the runtime of the algorithm. FP-Growth algorithm decreases the number of dataset scanning contrasted to Apriori, but expend more memory when it operates on large dataset instances.

VII. CONCLUSION AND FUTURE WORK

Due to huge advancement of web-based applications, most of the research is happening in web usage mining. It considers interpreting the look through behavior of website users and employing the perceived knowledge to elevate perfection of browsing experience. This paper presented an outline of all the aspects of web usage mining method. This paper described the recent approaches for preprocessing of web server log. This work also highlighted techniques and algorithms used in pattern discovery and analysis. Present work identified gaps that are present in several existing preprocessing and pattern discovery algorithms. Study and analysis of frequent item set mining algorithm reveals that for small instances of dataset, FP-growth, Apriori algorithm is better than SS-FIM. However, SS-FIM performs well for dataset with large number of instances.

Scalability of preprocessing algorithm for web usage mining is the focus of the future work. Design and implementation of efficient hybrid method in pattern discovery and analysis phase to extract more knowledge from web log file.

ACKNOWLEDGMENT

We are greatly thankful to our respective departments for providing excellent lab facilities that made this work possible.

REFERENCES

[1] P. Dhanalakshmi, K. Ramani, B. EswaraReddy, "*The Research of Preprocessing and Pattern Discovery Techniques on Web Log Files*", IEEE 6th International Conference on Advanced Computing (IACC), pp.139 – 145, 2016.

[2] P. Sukumar, L. Robert, S. Yuvraj, "*Review on Modern Data Preprocessing Techniques in Web Usage Mining*", International Conference on Computational System and Information Systems for Sustainable Solutions, 2016.

[3] Mitali Srivastava, Rakhi Garg, and P K Mishra, "*Analysis of Data Extraction and Data Cleaning in Web Usage Mining*", ICARCSET - 15, March 06 - 07, 2015, Unnao, India.

[4] Neha Goel, C. K. Jha, "*Preprocessing web logs: A critical phase in web usage mining*", International Conference on Advances in Computer Engineering and Applications, pp. 672 – 676, 2015.

[5] B.UmaMaheswari, P.Sumathi,"A New Clustering and Preprocessing for Web Log Mining", World Congress on Computing and Communication Technologies, 2014, pp.25 – 29.

[6] Nisarg Pathak , Viral shah, Chandramohan Ajmeera, "*A Memory Efficient Algorithm with Enhance Preprocessing Technique for Web Usage Mining*", ICTCS'14 November 14 - 16 2014, Udaipur, Rajasthan, India.

[7] K. Sudheer Reddy, G. Partha Saradhi Varma, and M. Kantha Reddy, "*An Effective Preprocessing Method for Web Usage Mining*", International Journal of Computer Theory and Engineering, Vol. 6, No. 5, October 2014.

[8] Priyanka S. Panchal ,Prof. Urmi D. Agravat, "*Hybrid Technique for User's Web Page Access Prediction based on Markov Model*", IEEE - 31661, 4th ICCCNT 2013, July 4-6, Tiruchengode, India.

[9] P. Nithya, P. Sumathi, "*An Enhanced Preprocessing Technique for Web Log Mining by Removing Web Robots*", IEEE International Conference on Computational Intelligence and Computing Research, pp.1-4, 2012.

[10] K.Sudheer Reddy, Dr. G. Partha Saradhi Varma, Dr. I. Ramesh Babu, "*Preprocessing the Web Server Logs – An illustrative approach for effective usage mining*", ACM SIGSOFT Software Engineering Notes, Volume 37, May 2012.

[11] D. Kerana, Kaliyamurthie K. P. "*Mining Frequent Item sets in Association Rule Mining Using Improved SETM Algorithm*", Artificial Intelligence and Evolutionary Computations in Engineering Systems, Advances in Intelligent Systems and Computing, vol 394. Springer, New Delhi, India 2016.

[12] Yakhchi S., Ghafari S.M., Tjortjis C., Fazeli M., "*ARMICA-Improved: A New Approach for Association Rule Mining*", International Conference on Knowledge Science, Engineering and Management, Springer, pp 296-306, July 2017.

[13] Rashmi V. Mane, V.R. Ghorpade, "*Predicting Student Admission Decisions by Association Rule Mining with Pattern Growth Approach*", International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT), Dec 2016 , Mysuru, India.

[14] Janisa Colaco, Jayashri Mittal, "*Improvised Hybrid model for User Navigation Pattern Prediction*", International Conference on Communication and Electronics Systems (ICCES), 2016.

[15] S. Vijayarani, S. Sharmila, "*Comparative Analysis of Association Rule Mining Algorithms*", International Conference on Inventive Computation Technologies (ICICT), 26 - 27 August 2016, Coimbatore, India.

[16] Aashna Agarwal, Nirali Nanavati, "*Association rule mining using hybrid GA-PSO for multi-objective optimization*", International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, Chennai, India.

[17] Deepti Sahu, Rishi Soni, "*A New Method for detecting User Behavior from Web Server logs*", International Conference on Computational Intelligence and Communication Networks, December 2015, Jabalpur, India.

[18] Manju, Chander Kant, "*Mining Association Rules Directly Using ACO without Generating Frequent Item sets*", International Conference on Energy Systems and Applications, November 2015, Pune, India.

[19] Ashika Gupta, Rakhi Arora, "*Web Usage Mining Using Improved Frequent Pattern Tree Algorithms*", International Conference on Issues and Challenges on Intelligent Computing Techniques, April 2014, Ghaziabad, India.

[20] Zailani Abdullah, Tutut Herawan, Noraziah Ahmad, "*Mining Indirect Least Association Rule from Students 's Examination Datasets*", ICCSA 2014, pp. 783–797.

[21] Avadh Kishor Singh, Ajeet Kumar, "*Association Rule Mining for Web Usage Data to Improve Websites*", International Conference on Advances in Engineering & Technology Research, August 2014, Unnao, India.

[22] Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley,Reda Alhajj, "*Effective web log mining and online navigational pattern prediction*", Knowledge-Based Systems Volume 49, September 2013, pp. 50-62.

[23] http://kdd.ics.uci.edu/databases/msnbc/msnbc.html