Analysis of Web Log from Database System utilizing E-web Miner Algorithm

Mukul B. Chavan
P.G. Student, Dept. of Computer Engineering
G.H. Raisoni College of Engg & Mgmt, Wagholi
Pune, India
mukulchavan09@gmail.com

Mrs. Sarita Patil
Faculty, Dept. of Computer Engineering
G.H. Raisoni College of Engg & Mgmt, Wagholi
Pune, India
Sarita.patil@raisoni.net

Abstract— Enormous content of information on the World Wide Web makes it clear for contender for data mining research. Data Mining Technique application is used to the World Wide Web referred as Web mining where this term has been used diverse ways. Web Log Mining is one of the Web based application where it confronts with large amount of log information. In order to produce the web log through portal usage patterns and user behaviors' recognition, this intended work is an endeavor to apply an efficient web mining algorithm for web log analysis, which is applied to identify the context related with the web design of an e- business web portal that requests security. Because of tremendous utilization of web, web log documents have a tendency to become large resulting in noisy data files. It can find the browsing patterns of client and some class of relationships between the web pages. Here we analyze the logs using web mining algorithm. Whatever the result we will get compare within the Apriori, AprioriAll and E-Web Miner Algorithm. Through the analysis we recognize that web mining algorithm called E-web miner i.e. Efficient Web Mining performs better considering space and time complexity. It can likewise be verified by comparison, candidate sets are much smaller and our results show number of database scanning get reduced due to implementation of E-Web Miner Algorithm.

Keywords- *Web Mining, Web Log Analysis, Server Log File, E-Web Miner, Apriori Algorithm, AprioriAll Algorithm.*

*****

I. INTRODUCTION

Data mining is a technique used to find useful and significant information to guide professional decisions and other scientific research. Predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. It is a cost-effective way of analysing large amounts of information, especially when a human could not analyse such datasets. Enormous utilization the web has made automatic knowledge extraction from Web log documents a need. Web mining is an application of data mining field, which extract interesting and potentially useful patterns and hidden information from web documents and web activities. This can enhance the effectiveness of their Web sites by adapting the information structure of the sites to the users' behaviour. In web mining information is in unstructured configuration as Log documents [1]. Log files are some kind of machine readable semantics. Web Mining can be comprehensively divided into three different classifications as per the sorts of data to be mined.

A.Web Content Mining: It is the process of extracting valuable data from the contents of Web. Content information corresponds to assembling of facts a Web page was designed to convey to the users. It may comprise of structured, unstructured information like text, pictures, sound, video, E-mails, HTML files or organized records for example lists and tables.

B. Web Structure Mining:

Web structure mining exploits the additional hidden information that is often contained in the structure of hypertext. The structure of Web graph comprises of Web pages as nodes, and hyperlinks as edges linking between two related pages. It can be viewed as the methodology of finding structure data from the Web. This sort of mining can be further separated into two types based on the kind of structural information utilized. It may comprise of Hyperlinks, Document Structure.

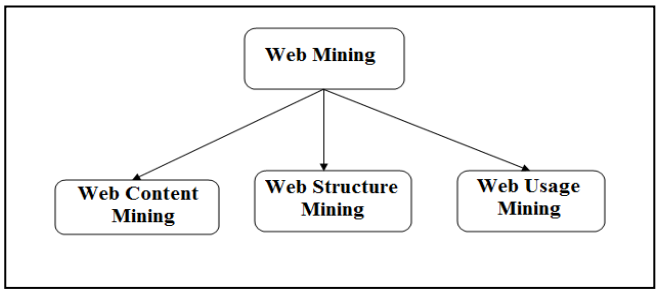

Figure 1. Web Mining Taxanomy.

C. Web Usage Mining: It is the use of data mining techniques to find out interesting usage patterns from Web information, to understand and serve the requirements of Web based applications. Usage data captures the identity of Web users and browsing behavior at a Web site. Catching, Modeling and analyzing of behavioral patterns of users are the objective of this web mining category. Web usage mining procedure can be separated into three independent tasks: Pre-processing, Pattern discovery and pattern analysis. Web Usage Mining mines the secondary data like Web server access logs, browser logs, client profiles, registration data, client sessions or transactions, cookies, user queries, derived from the interactions of the users during certain time of Web sessions [5]. Web Usage Mining understands the user behaviour in interacting with a

particular web site. Web usage mining uses web logs to record user access patterns. Log files are created by web servers and filled with information about user requests on a particular Web site. We make an endeavor to apply an efficient web mining algorithm for web log analysis. The results acquired may be utilized against class of problems; from search engines in order to identify the context on the basis of association to web site design of an ecommerce web portal that demands security. This work intends to illustrate that the E-Web Miner has much better execution in provisions of time and space complexity than Apriori and AprioriAll Algorithm and affirms the appropriateness of result acquired by giving a trace back route for candidate set pruning for the algorithms. The number of data base scanning's significantly gets decreased in proposed algorithm and the candidate sets are discovered to be much minor.

II. LITERATURE SURVEY

Various algorithms have been developed in recent years for web log analysis according to client and server perspective. Apriori algorithm introduced by Dunham in 2003 then some progressions are permitted to sorting of information according to Userid and Timestamp which available in log data base related to each user worked with WWW. Basic differentiation among Apriori and AprioriAll is that AprioriAll algorithm utilizes full join of candidate set and in case of Apriori forth join is used so AprioriAll is more used in usage mining.

P.Lopes, B.Roy proposed a system providing real time dynamic recommendation to all visitors of the website independent of been enlisted or unregistered [2].Action based rational recommendation technique is proposed that uses lexical patterns to generate item recommendation. H. Sha, Qin, Sun, Liu[3] in their paper focus on the data cleaning problem and proposed a new filtering method EPLog Cleaner in order to further improve data quality by removing specific irrelevant requests which share the same time characteristic from proxy log. In our approach, we confirmed that such requests are much more regularly and even periodically than others, and this can be used as the characteristic in distinguishing them from others. By using such characteristics, researchers may further reduce irrelevant items in proxy log while still ensure the remaining items are valid.

Wang Tong HE Pi-lian [4], exhibited that the probability and importance about applying Data Mining in Web log mining and demonstrated few issues in the routine searching engines. After that it offers an improved algorithm based on AprioriAll algorithm, which has been used in Weblogs mining widely. Most system outcomes show the improved algorithm has a less complexity of time and space.

Wen-Hai Gao acquainted the strategy how to utilize data mining technique to excavate the client behavior pattern from Web log, and play an importance on analyzing client behavior pattern recognition system and its application [6]. In this article, they analyzed a particular case in which operationalized website-based variables to understand R&D activity, amongst others [7]. The processes of data retrieval, preparation, cleaning and analysis for web content data are more complex compared to conventional data sources. Care in the interpretation of the website data is particularly important. Approaches to searching that are too simplistic and overlook context can lead to false positives or incorrectly dropping instances.

A.V. Krishna Prasad and Dr. S. Ramakrishna, details regarding an experience utilizing open Web Application Programming Interfaces (APIs) that has been made accessible by major Internet companies in a project to teach Web applications [8]. The observations of the performance and a survey of the APIs show that we achieve project objectives and acquire valuable experience in leveraging the APIs to build interesting Web applications. B. Mobster, R. Cooley, J. Srivastava [9] illustrates a methodology to usage-based Web personalization considering the full spectrum of Web mining techniques and activities. Approach heavily utilizes data mining techniques, thus making the personalization process both automatic and dynamic. [10] Grace, V.Maheswari, D. Nagamalai in their work gives a detailed look about log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that can be used in the log files which in turn give way to an effective mining. It also provides the idea of creating an extended log file and learning the user behaviour. U. Patil, S. Pardeshi, presented paper that underline on user future next request prediction using web log record, click streams record and user information. The aim is to offer past, current assessment and update in web usage mining- future request prediction [11]. This paper presents the Prediction of User navigation patterns using Clustering and Classification (PUCC) from web log data. [12] In the first stage PUCC concentrates on isolating the potential clients in web log data, and in the second stage clustering process is used to group the potential clients with comparable interest and in the third stage the results of classification and clustering is utilized to foresee the client future solicitations.

III. PROPOSED SYSTEM

Web log analysis software (also known as web log analyser) is a type of Web Analytic software that parses a server log file from a web server, and taking into account the values contained in the log file, derives indicators about when, how, and by whom a web system is visited. Usually reports are generated from the log files immediately, but the log files can alternatively be parsed to a database and reports generated on demand.

 ISSN: 2321-8169
4730 - 4734

This system is an attempt to apply an efficient web mining algorithm for web log analysis. The results obtained from the web log analysis may be applied to different problems of search engines in order to identify the context and to design web site of an e-commerce as per the user's behavior. Generally, the main objective of this project is to Web Usage Mining process, specifically:

- To preprocess server logs files from the Web servers for determining and discovering the user access pattern.
- To apply algorithms and analyze the outputs usage patterns and user behaviors from the Web Usage Mining implementation process.
- To reduce number of database scanning by implementation of E-Web Miner Algorithm and decrease candidate sets size.

System Overview

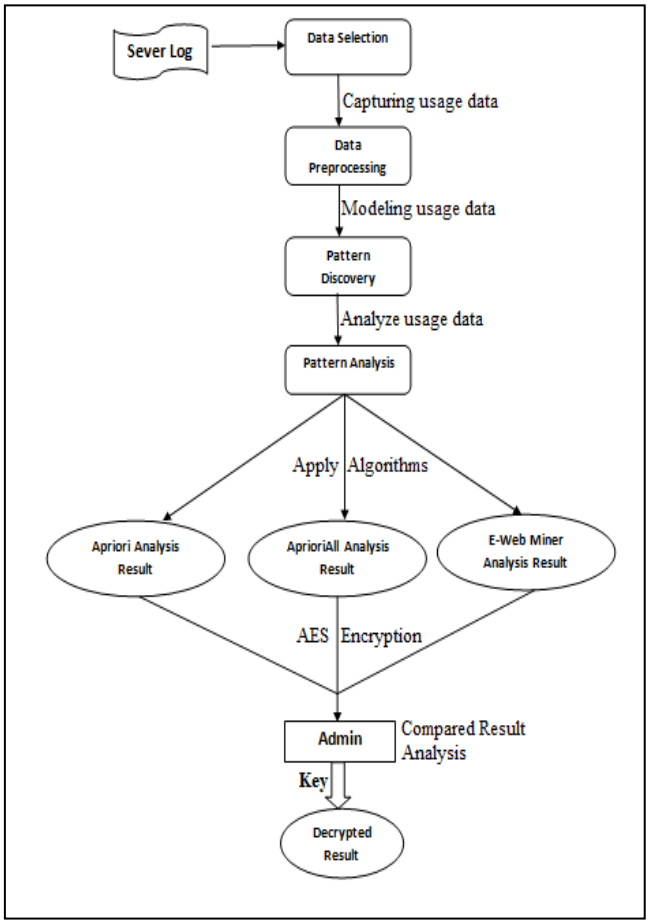

Fig 2: System Overview

Server Log File: The server log files are retrieved from the server. Web log file is automatically created and kept by a web server. Every "hit" to the Web site, including every observation of a HTML document, image or other object, is logged. The unprocessed web log file format is essentially one line of text for each hit to the web site. Thus server log file are been chosen for further analysis. The information in web log file keep up a correspondence to the access patterns of different user's of the overall web traffic, ranging from single-user, single-site browsing behaviour to multi-user, multi-site access patterns. A log file record contains essential information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL [12], and the browser information.

Data Selection: In the data selection phase, selection of the server log files must be done carefully; there are several facilities such as My Portfolio and Resources. The server log file also includes the mix of log file for every transaction between the facilities in the Web portal.

Data Preprocessing: Data Pre-processing comprises of all the considerable actions taken before the actual pattern Analysis phase process starts. The pre-processing steps include cleaning, user identification and session identification. Cleaning is the process which removes all entries which will have no use during analysis or mining. The major task in this phase is includes handling missing values, identifying outliers, smooth out noisy data and correct inconsistent data. The large amount of server log file data becomes the most challenging problem to handle during the Data Preprocessing phase. We use Clean and Sort in our implementation. It gives total no of users in Particular Log.

Pattern Discovery: The pattern discovery has three major operations of concern, Association (i.e. which pages to be accessed collectively), Clustering (i.e. finding groups of users, transactions, pages, etc.), and Sequential analysis (the order in which web pages tend to be accessed). Here all the algorithms are applied to the preprocessed data.

In session wise view we undergo following type of log format

User Agent: This is nothing but the IP address from where the user sends the request to the web server.

Date: The time or date when the user surf web page from the web site. This is identified as the session.

URL: The resource accessed by the user. It may be an HTML page, a CGI program, or a script.

Request type: The method used for information transfer is noted. The methods like GET, POST.

Response Type:

100 HTTP_INFO
200 HTTP_SUCCESS
300 HTTP_REDIRECT
400 HTTP_CLIENT ERROR
500 HTTP_SERVER ERROR

Pattern Analysis: Pattern analysis the inspiration is to filter out uninteresting rules or patterns found in the earlier phase. During the Pattern Analysis phase, the descriptive method is being used to analyse the data after the various algorithm

implementations such as general summary of the Web usage and customer behaviours. The analysis also tries to find out the top visitors for each facility or option that being provided by the portal. Beside the option analysis, the sever log files also trace the information of documents that was downloaded.

Applying Algorithms: Apriori, AprioriAll and E-Web Miner Algorithm are applied individually to similar log file and then results are calculated with respect to time and Log records. It also describes differences between all algorithms in terms of time and space complexity.

Security: Here we are going secure the log analysis result using AES. In this module the analysed data will be gathered in encrypted format so its not be attacked or reveal by any attacker. This data can be viewed by only authenticated person by the authentication process such as Admin who holds the desired decryption Key.

AES Public key Encryption Algorithm:

The Advanced Encryption Standard (AES), is a specification for the encryption of electronic data established by the U.S. National Institute of Standards and Technology in 2001. AES is based on the Rijndael cipher developed by two Belgian cryptographers, Joan Daemen and Vincent Rijmen, who submitted a proposal to NIST during the AES selection process. Rijndael is a family of ciphers with different key and block sizes. For AES, NIST selected three members of the Rijndael family, each with a block size of 128 bits, but three different key lengths: 128, 192 and 256 bits.

AES has been adopted by the U.S. government and is now used worldwide. It supersedes the Data Encryption Standard (DES), which was published in 1977. The algorithm described by AES is a symmetric-key algorithm, meaning the same key is used for both encrypting and decrypting the data. The block cipher Rijndael is designed to use only simple whole-byte operations. Also, it provides extra flexibility over that required of an AES candidate, in that both the key size and the block size may be chosen to be any of 128, 192, or 256 bits. During an early stage of the AES process, a draft version of the requirements would have required each algorithm to have three versions, with both the key and block sizes equal to each of 128, 192, and 256 bits. This was later changed to make the three required versions have those three key sizes, but only a block size of 128 bits, which is more easily accommodated by many types of block cipher design. In our system we implement it for the result generation phase of analysis. Only the desired user will be able to view the analysis.

Results: As stated above, this implementation will focus on Web Usage Mining of Portal. The results of this study are divided into two areas where the first section will discuss about the general descriptions of the access pattern and users behaviours of Portal (descriptive statistic). Another section will display the supports and confidences of the different level in Portal. All the results will display using certain chart such as graphs and tabular data result to make it easier understand.

Proposed Algorithm

E-Web Miner is the improvisation of Web mining algorithms which removes the loopholes in the Apriori-All Algorithm. Following are the steps of the algorithm.

1. Make the set of web pages in the ascending order for the various users.
2. Now assign the set of pages in the string array 'a' for the user 'u'.
3. Initialize the f=0, max= 0, where f is frequency and max is maximum.
4. Consider I vary from 1 to n; also J varies from 0 to (n-1)
5. If substring (a [I] ; a [J])
   f=f+1;
   END IF
   b [I] = f;
   If max ≤ f
   Max = f;
   END IF
   6. Find out the positions in array b, where the value is nearly equal to maximum value and choose the corresponding substring from array 'a'.
   7. Repeat the step 6 and produce output for all the substring with their positions, which is intended output.

Mathematical Model

1. Calculating the set of web user
   WU = {wu1, wu2, wu3…}
   Where 'WU' is main set of Web user like wu1, wu2, wu3…..
2. Identify the set of logs visited
   V F = {vf1, vf2, vf3…}
   Were 'VF' is main set of logs visited like vf1, vf2, vf3. .
3. Calculating the set of logs after Pattern Analysis
   AF = {af1, af2, af3…}
   Where 'AF is main set of logs after Pattern Analysis like af1, af2, af3. . . .
4. Calculating the set of logs after Pattern Discovery using Apriori
   UF = {uf1, uf2, uf3….}
   Where 'UF' is main set of logs after Pattern Discovery using Apriori like uf1, uf2, uf3. . .
5. Calculating the set of logs after Pattern Discovery using AprioriAll.

FE = {fe1, fe2, fe3…}

Where 'FE' is main set of logs after Pattern Discovery using AprioriAll like fe1, fef2, fe3…..

6. Calculating the set of logs after Pattern Discovery using E-Web Miner.

M = {m1, m2, m3…}

Where 'M' is main set of logs after Pattern Discovery using E-Web Miner like m1, m2, m3. .

IV. EXPRIMENTAL RESULTS

A. *Dataset*

In this work we have used Server Log file as input to system.

B. *Results*

As per this proposed System, we can achieve various Performances for log analyzer like Time and space Efficient Web mining Algorithm, Free up User Behavior, Secure result transmission.

From below graph, it proves proposed E-Web miner system performs better than existing Apriori, AprioriAll system in which time required for analyzing log. Which proves that proposed E-web Miner has lower time complexity as compared to rest of mining algorithms.

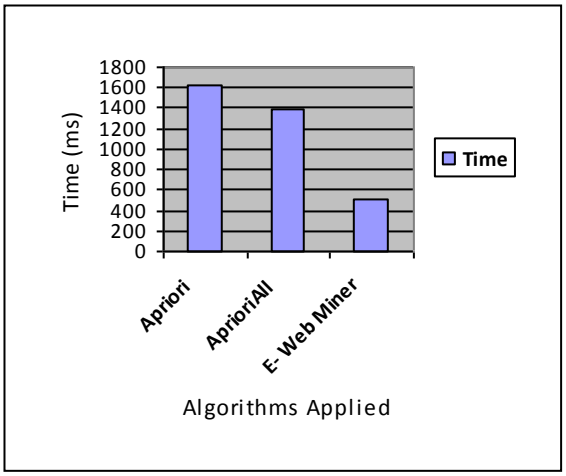

Fig 3. Comparison Graph of various Algorithms implemented.

V. CONCLUSION AND FUTURE WORK

Proposed procedure for programmed web log data mining by E-Web Miner algorithm is more effective. It stands above other web mining algorithms. The Proposed new Web Mining Service can avail private log file of different website to general public. This log information will be helpful for E-commerce. E-Web Miner may take into consideration the support and the confidence of any sequential pattern of web pages of users. This may give further refinement in the result of candidate set pruning. The number of data base scanning's drastically gets reduced in E-Web Miner and the candidate sets are found to be much smaller in stage wise comparison with Apriori, AprioriAll Algorithm.

Future work is to carry out path analysis on this data. This analysis technique is utilized for determining the most often visited paths in a web site. We can make use of the user transactions, maximal forward transactions and user sessions to perform this analysis.

ACKNOWLEDGMENT

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We also thank the Computer Engineering Department of G.H. Raisoni College of Engineering and Management, wagholi, Pune authorities for providing the required infrastructure and support. Thanks to all those who helped me in completion of this work knowingly or unknowingly. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] Mahendra Pratap Yadav, Pankaj Kumar Keserwani and Shefalika Ghosh Samaddar, "An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner", RAIT-2012 ,978-1-4577-0697-4/12/ IEEE 2012.

[2] P.Lopes,B.Roy, " Dynamic recommendation using Web usage mining for E-Commerce users", ICACTA Elsevier 2015

[3] Hongzhou Sha, Tingwen Liu, Peng Qin, Yong Sun, Qingyun Liu,"EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining", ITQM , Elsevier 2013.

[4] Tong, Wang and Pi-lian, "Web Log Mining by an Improved AprioriAll Algorithm", World Academy of Science, Engineering and Technology, Vol 4 2005 pp 97-100.

[5] Sachin Pardeshi and Tareek Pattewar, "Free User's Behaviour Information from Central Database System (Web Mining)", Proceedings of 7th International Conference on Intelligent Systems and Control (ISCO 2013)978- 1- 4673-4603-0/12/c 2012 IEEE.

[6] Wen-Hai Gao, "Research on Client Behaviour Pattern Recognition System Based On Web Log Mining", Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, IEEE 2010.

[7] Abdullah Gok , Alec Waterworth , Philip Shapira, "Use of web mining in studying innovation", Scientometrics, Springer 2014.

[8] A.V. Krishna Prasad, Dr. S. Ramakrishna, "Retrieving business applications using Open Web APIs Web Mining dashboard application case study," (UCSIT) International Journal of Computer Science and Information Technologies, Vol. I (3), 2010.

[9] Bam shad Mobster, Robert Cooley, Jaideep Srivastava "Automatic Personalization Based on Web Usage Mining" ACM, Volume 43 Issues 8, Aug 2000.

[10] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai , "Analysis of Web Logs and Web user in Web Mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.

[11] U.M. Patil and S.N .Pardeshi. "A survey on user future request prediction: Web Usage Mining" (UETAE) International Journal of Emerging Technology and Advanced Engineering, Vol.2, pp. 121-124 March 2012.

[12] V. Sujatha, Punithavalli, "Improved User Navigation Pattern Prediction Technique from Web Log Data", International Conference on Communication Technology and System Design, Elsevier 2011.