IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Preprocessing Application based on Structured Query Language for Web Log Mining

K. Vadivazhagan[1*], M. Karthikeyan[2]

[1, 2]Division of Computer and Information Science, Annamalai University, Tamilnadu, India

[*]*Corresponding Author: vadivazhagan.k@gmail.com, Tel.: +91-9994580854*

DOI: https://doi.org/10.26438/ijcse/v7i3.544549 | Available online at: www.ijcseonline.org

Accepted: 14/Mar/2019, Published: 31/Mar/2019

***Abstract*—** Web Log Mining, also known as Web Usage Mining (WUM) is the application of Data Mining techniques, which is applied on web log data to extract interesting patterns. An enormous increase in the use of web applications as medium of the organizations and institutions, the web page hits are consistently increasing. The web servers have the facility to save the web navigational sequence as web log file. The enormous amount of irrelevant information in the web log file demands proper preprocessing. This renders the file, with the intent of making it more appropriate for a variety of downstream purposes such as analytics. There are various traditional techniques involved in preprocessing. The implementation of preprocessing model presented in this paper over other traditional preprocessing methods is to employ an efficient Structured Query Language (SQL) based technique. The proposed SQL based preprocessing technique reduces process time drastically. The resulting structured log file is well suited for further pattern mining and analytics.

***Keywords*—** Preprocessing, Web Log Mining, Server log, User Identification, Session Identification

I. INTRODUCTION

Web log preprocessing and cleansing are becoming more and more important in today's analytics. During the process of constructing the analytical model using various techniques or Machine Learning, the data set is collected from various resources such as a server log file and database.

The collected data cannot be used directly to perform analytical process [1]. To overcome this issue, the data has to be prepared properly. It includes two processes; they are (1) Data Preparation and (2) Data Preprocessing. Pre-processing in this context is the procedure of cleansing and preparation of web log data which is to be mined. It is a fact that unstructured web log data on the server log file, contain significant amounts of noise. By the term noise, we mean data that do not contain any useful information for the analysis at hand.

Figure 1. Web application process

The list of requests log or errors log which have succeeded or failed logs are collected in web server log file. The client request and server response states depicted the above Figure 1. It records not only for the web page which has links to a request file that exists records as success log parameter values and does not exist records as error log parameter values such as IP address of the client, referred and referral URL, status code and user agent of the client where requested web pages. As well the user who is not having permission to access a page, the user request may fail also recorded.

A. Data Preparation

Data Preparation is an important part of Web Log Preprocessing. It includes two concepts such as (1) Import Data to Database and (2) Formatting Data Structure. These two processes are necessary for achieving better precision and performance in the Machine Learning and Deep Learning analytics. Figure 2 describes various steps involved in preprocessing.

- Data Preparation is a preprocessing step in which web log data from web server is cleaned and transformed to improve its quality prior to its use in business analytics.

- The web log data collected from web server has to be uploaded to the database.
- The database then has to be restructured according to the type of data. It can involve changing the formats of dates.

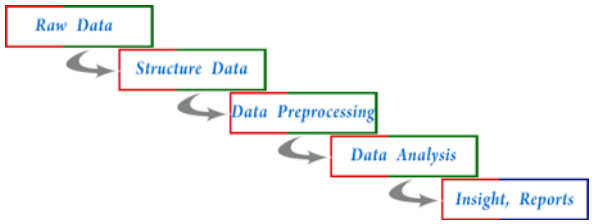

Figure 2. Data Preparation and its process

B. Data Preprocessing

When data is of worthy condition it can be properly processed and analyzed, leading to insights that help the organization make better analysis [2]. High-quality data is essential to business intelligence efforts and other types of data analytics, as well as better overall operational efficiency.

Data preprocessing is done using database-driven applications such as SQL based applications. During preprocessing data subjected to a series of steps and they are:

- Data Cleansing: Data is cleansed through cluster processes.
- Data Integration: Data with different groups are put together.
- Data Transformation: Normalize data according to requirements of analysis.
- Data Reduction: Reduced data can be easily processed.

In this study, the web log file is subjected to data preprocessing. The restructured file resulting from preprocessing is used for further analysis. The SQL based techniques are employed in preprocessing of the data. This enables one to manage a website effectively to satisfy the needs of the user [3],[4].

II. LITERATURE REVIEW

A survey of literature is carried out on "Various techniques involved in web log preprocessing".

Web log mining is mainly related to web usage mining [5]. The various data preprocessing techniques are used to determine the efficiency of the algorithms, its precincts, and its stands are verified. Various preprocessing algorithms and its heuristic techniques are applied to examine by using programming languages. The preprocessing algorithms are used to parse the raw log files that involve splitting number of log files. They are then cleansed to obtain higher quality of data. Based on this preprocessed data, the unique users can be identified which in turn helps to identify user sessions.

The process of web usage mining is implemented to discover the user patterns from the web log data collected from the web server [6]. After analyzing this data, the user behaviour can be predicted and accordingly sequence of the web can be restructured and reordered. A custom algorithm is designed for the Clustering process which was aimed to provide more efficient algorithm and results are compared to the various Clustering algorithms.

Information collected from web server logs of web site can be very useful for purposes such as research and technical analyses, the design of web sites, applications and the optimization of their functioning [7]. The analysis of web server logs, or web log mining, are usually processed in three main phases and they are preprocessing, pattern discovery and pattern analysis.

A new web log mining method was proposed to determine web access procedure from manipulated web usage logs which integrates data on user behaviours through communication tracking [8]. The raw data have to be preprocessed in order to improve the quality of the data. The significance of information preprocessing method was discussed in receiving the essential information effectively. This preprocessing technique was used to process and analyze the web log data for extraction of user navigational patterns. The algorithm fuzzy association rule mining decrease user exploration to derive decision making. The proposed method removes the unwanted records from the web log and developed Personalized Ontology, assists various semantic web applications. Finally, the efficiency of this method was illustrated by the experimental results in the framework of data pre-processing and cleansing extraneous data for personalized ontology.

To identify the user navigational behavior using web usage mining, the web logs play a vital role [9]. While many pattern mining methods are used to identify user behavior, the accuracy & quality of pattern mining algorithms can be improved with the help of preprocessing techniques. Various activities like identifying the number of unique users, reducing the size of log file, identifying the sessions are done with the help of preprocessing techniques in the existing algorithms. A new algorithm named Enhanced User Behavior (EUB) which was proposed to identify user and groups. It brings into discussion about the concepts of web

log preprocessing, and various clustering preprocessing techniques.

Suneetha K.R, R.Krishnamoorthi discussed an algorithm for Data cleansing, user identification, and session identification [10]. The algorithm is applied on web log data in order to reduce the size of the actual data, to obtain the unique users and to find the navigational actions of the user using the session, with start page time and end page time. They also employed a new method to access the usage pattern of preprocessed data where the results of preprocessed web server logs were stored using snow flake schema of data warehouse to smooth the progress of easy retrieval and analysis.

III. SQL BASED PREPROCESSING TECNIQUES

The web log preprocessing is a method for the purpose of explanatory data summarization which includes data cleansing, data integration, data transformation, data reduction and data discretization. Data collected in the real world is unstructured like semi structured, noisy, and inconsistent. To ensure the quality data and quality mining results and to find accurate knowledge from the available data it should properly cleansed before mined. For example, replicated or missing data may be caused incorrect or even deceptive statistics. Also the data repository needs consistent integration of quality data. The data extraction, cleansing, and transformation contain the mainstream work to build a data warehouse. It leads to preprocess as an important process in the web log mining. The web logs collected from the web server which is a hyper text transfer protocol request from the user and it passes them to the web server. Web logs for the particular user are maintained in the browser machine. Browsers are programmed and scripting languages are employed in it to collect the client side data. There are three types of web log formats are available [11], [12]. The formats are defined by World Wide Web consortium (W3C), which is an extended default log file format and includes list of agents visiting the websites, duration of session, time of visit, page sequences, IP address of client system and status code success/error. It is difficult to mine the web logs directly and therefore knowledge mining algorithms are used to extract the features. To such extent the preprocessing techniques are inevitable to make data consistent and complete. The following Figure 3 depicts the various steps involved in data preprocessing.

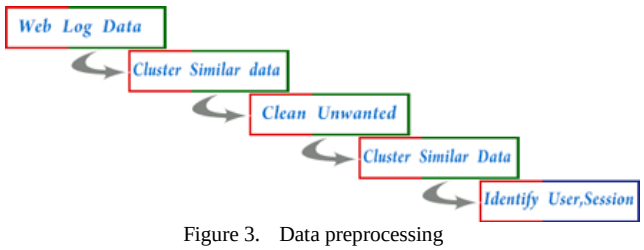

Figure 3. Data preprocessing

As stated, preprocessing is an important process in web log mining and plays a key to successful mining [7]. It's a technique to remove the irrelevant information from the web logs and smooth the process of the effective pattern mining. The steps involved in data preprocessing consists of data cleansing, user identification and session identification.

A. Data cleansing

Data cleansing is the process of removing irrelevant records that are not necessary for mining [13]. Data cleaning includes, (1) Removal of records of graphics, videos and the format information. (2) Removal of records with the failed HTTP status code.

B. User and Session Identification

The process of user identification is to find out the unique user, based on their IP address of request machine and its user agent from the browser properties stored in the web log [14]. In this work, sessions of the user are separated using threshold time which is normally 30 minutes as a single user. Web log preprocessing facilitates to remove unwanted click streams from the log file and it reduces the size of original log file by more than 75%.

IV. PROPOSED METHODOLOGY APPLIED IN PREPROCESSING

The proposed methodology concentrates highly on SQL based approaches in preprocessing, while the existing system uses various traditional preprocessing techniques based on scripting language. The proposed system makes use of partition methods as SQL views to group the similar kind of data to be separated in to different virtual table. Instead of removing the irrelevant data from table, the table views are created so that the irrelevant logs may be used for further analysis. To such extent the relevant information is stored in another SQL view as virtual table. There after the newly created SQL view can be used for all other analysis instead of using the main table.

A VIEW is a virtual table, through which a selective portion of the data from main log table have to be separated in two views as one for noisy log and other for cleansed log. Views do not contain data of their own. They are used to restrict access to the database or to hide data complexity. Views can provide advantages over tables. Views can represent a subset of the data contained in a table. So this approach can provide faster results than scripting language. The proposed SQL based algorithm considers the following features to preprocess the log file, identify the user and to identify the sessions. The web log date is collected from www.annamalaiuniversity.ac.in web server and it is limited

to one lakhs for preprocessing. Table 1 shows the single log entry of web log data [15].

Table 1. Sample Common Log Format from Log File

64.233.173.133 - - [05/Mar/2019:17:43:38 +0530] GET /index.php HTTP/1.1 200 53537 http://www.annamalaiuniversity.ac.in/dde/ Mozilla/5.0 (Linux; Android 6.0; Lenovo A7020a48) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/72.0.3626.105 Mobile Safari/537.36

- During the data cleansing process, explicit relevant and irrelevant information are considered and the irrelevant requests are not removed from web logs.
- Users are identified using their IP address and user agent of browser properties.
- User session is calculated based on the threshold time spent on website by a particular user.
- Frequency value is presented based on the number of navigational sequence created be the user on the web site.

The applied preprocess technique also finds the status codes of the logs and accordingly it is shown in the Figure 4. In this, the codes 200 and 206 denotes the successfully visited pages of 1,00,000 log entries while other 3 status codes denotes failure visits for various reasons. Table 2 depicts the summary statistics of post preprocessed log details.

Table 2. Summary Statistics of Preprocessed Log Details

Status Code	***Count***	***Percentage %***
200	90832	91.3
206	1657	1.67
301	499	0.5
304	3145	3.16
404	3356	3.37
Total Status Codes	99489	100

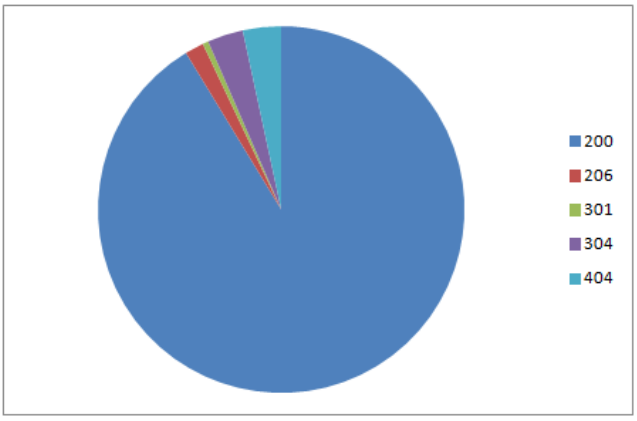

Figure 4. Distribution of status codes of web log

The pseudo code for preprocessing is shown in Table 3. Moreover, this approach is not specific to a noise removal of log file. It is also used to store the various properties of the log file details. The method can be used for big data related processing.

Table 3. SQL Based Algorithm

SQL Based Algorithm
Input: Web_log_file_txt
Output: Refined_web_log_file_table
Begin
1. Create table with appropriate structures
2. Create indexing
3. Alter table to remove meaningless column
4. Change date column data type
5. Create 2 SQL Views for cleansed log and noisy log using select queries with condition clause
6. Call group_process to group user
7. Call session_process to find number of sessions
8. End the process

In the proposed SQL based algorithm, the density of the web log contents reduces considerably and it makes use of the data appropriately for the knowledge mining and

visualization. Following Table 4 show the properties of the preprocessing values in numerals and its parallel percentage values.

Table 4. Preprocessed Web Log Details

Description	*Log Count*	*Bytes (KB)*	*Percentate %*
Logs uploaded	1,00,000	63,18,294.62	100
Logs after cleansing	14,844	22,44,859.43	14.84
Logs cleansed	85,156	40,73,435.2	85.16

From Figure 5, it is known that the number of file types identified by proposed algorithm. The two measures used are User identification and Session Identification. Session is calculated based on the threshold time spent on the website by a particular user. Jpeg files are found to be comparatively higher than other file types.

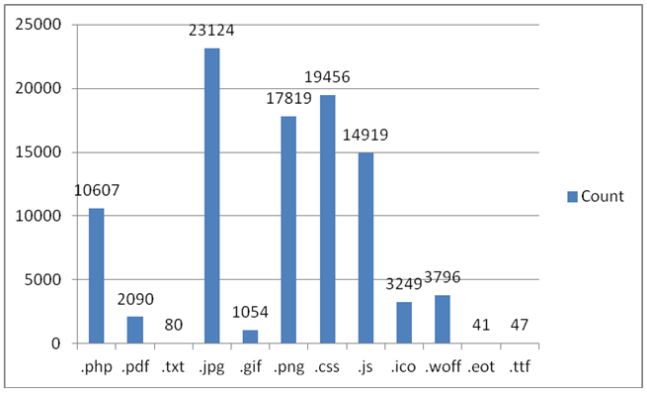

Figure 5. File type counts of web log

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

The web log files were downloaded from Annamalai University web server. The preprocessing algorithm is implemented by using PHP scripting language and MySQL as data base. It helps and supports the execution. SQL based cleansing algorithm is evaluated along with the current methods in terms of data cleansing, user identification, and session identification [15], [16], [17], [18].

Table 5. Details of Evaluation

Description	*Details*
Start Time	2019-03-05 17:39:48
End Time	2019-03-05 21:41:53
Total Time Spent	04:02:05
Unique IP	3,244
Max Used IP	1.186.37.93
Total Logs	1,00,000
After Cleansing Logs	14,844
Cleansed Logs	85,156
Total Users	4,103
Total Sessions	4,609
Threshold Time	30 Minutes
Average Time/Sessions	0:0:3
Total number of pages visited	12,509 pages
Average visit (Frequency)	3 pages

The Table 5 shows evaluation data for the data cleansing. A total of one lakhs log entries are taken for the preprocessing, 85,156 (85 %) logs marked as noisy log and are not suitable for mining. Only 14,844 (15 %) log is considered as cleansed log and it is more suitable for web log mining. User and session identification algorithms are used to identify the distinctive users and their sessions from the web logs. The cleansed logs are passed as input to the user and session identification phase. User and session identification algorithm results are shown below.

Thus, the proposed SQL Based algorithm proves its efficiency and effectiveness for finding the number of cleansed log and noisy log by using SQL View (Virtual Table techniques). SQL View is the process of partitioning cleansed log in to one view and noisy log onto other SQL View. The proposed algorithm takes an execution time of 0.39 Seconds to preprocess 843 logs while scripting language takes an execution time of 7.38 Seconds to preprocess same number of logs. For 10,000 logs the proposed algorithm takes 0.58 Seconds against 58.24 Seconds taken by scripting language. In same way for 1,00,000 logs the proposed algorithm takes 5.46 Seconds against more than 15 minutes taken by scripting language and its details shown in Table 6. However, with the increasing in the log data, the performance of time drops well when compared to scripting language. From the following details, SQL Based process significantly improves the performance of the application.

Table 6. WEB LOG PREPROCESSING PERFORMANCE

Log Count	*SQL Based*	*Scripting Language*
843	0.39 Seconds	7.38 Seconds
10,000	0.58 Seconds	58.24 Seconds
1,00,000	5.46 Seconds	15 minutes

VI. CONCLUSIONS AND FUTURE ENHANCEMENT

SQL Based Preprocessing application is developed to perform data cleansing, user identification and session identification. The implemented application preprocesses data effectively taken from web log files. The time consumption for execution of the application is lesser compared to other traditional preprocessing techniques. Many procedures and various methods are discussed for preprocessing the data collected from the web log files. If dealt with big data for preprocess, the proposed structured query language based application shows significantly improved performance.

In future direction, there is a huge connection between big data and data preprocessing. Using big data technologies and its various big data framework, such as R programming, Hadoop and Spark will emerge as new big data learning paradigms.

References

[1] J. Srivatsava, R. Cooley, M. Deshpande, and P. N. Tan, "*Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*", SIGKDD Explorations, Vol.**1**, Issue.**2**, pp.**12-23, 2000.**

[2] V. Chitraa, and Antony Selvadoss Devamani, "*A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing*", International Journal of Computer Applications, Vol.**34**, Issue.**9,** pp.**23-27, 2011.**

[3] K. Vadivazhagan and M. Karthikeyan, "*Preprocessing Techniques in Web Log Mining to Group Users and Identify User Session*", International Journal of Engineering Science Invention, Vol.**4**, pp.**26-33**, **2018.**

[4] K. Vadivazhagan and M. Karthikeyan, "*Mining Frequent Link Sets from Web Log Using Apriori Algorithm",* Journal of Computational and Theoretical Nanoscience, American Scientific Publishersber, Vol. **16**, pp**. 1–7, 2019**.

[5] P. Sukumar, L. Robert and S. Yuvaraj, *"Review on modern Data Preprocessing techniques in Web usage mining (WUM)"*, In the Proceedings of the 2016 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, pp.**64-69, 2016.**

[6] Bharat Chauhan, Hemant Kumar, Mihul Singh, Piyush Kumar and Sakshi Hooda, *"An Improved Preprocessing and Clustering Using Web Log Data"*, International Journal of Advanced Research in Computer and Communication Engineering, Vol.**5,** Issue.**11,** pp.**95–98, 2016.**

[7] Janusz Kacprzyk and Sławomir Zadrozny, *"Linguistic Summarization of the Contents of Web Server Logs via the Ordered Weighted Averaging (OWA) Operators"*, Fuzzy Sets and Systems, Elsevier North-Holland, Inc., Vol.**285**, pp.**182-198**, **2016.**

[8] F. Mary Harin Fernandez and R. Ponnusamy, "*Data Preprocessing and Cleansing in Web Log on Ontology for Enhanced Decision Making*", Indian Journal of Science and Technology, Vol.**9**, Issue.**10**, pp.**1-10, 2016.**

[9] S. Uma Maheswari and S. K. Srivatsa, "*An Application of Preprocessing and Clustering in Web Log Mining*", International Journal of Philosophies in Computer Science, Vol.**1**, Issue.**1**, pp.**21-30, 2015**.

[10] K. R. Suneetha, R. Krishnamoorthi, "*Identifying User Behavior by Analyzing Web Server Access Log File*", IJCSNS International Journal of Computer Science and Network Security, Vol.**9**, Issue.**4**, pp.**327-332**, **2009**.

[11] M. Udantha, S. Ranathunga and G. Dias, "*Modelling Website User Behaviors By Combining the EM and DBSCAN Algorithms*", In the Proceedings of the 2016 IEEE Moratuwa Engineering Research Conference (MERCon), Moratuwa, pp. **168-173, 2016**.

[12] Hsin-Jung Cheng and Akhil Kumar, "*Process Mining on Noisy Logs - Can Log Sanitization Help to Improve Performance?*", Decision Support Systems, Elsevier B.V., Vol.**79,** pp. **138-149, 2015**.

[13] Yin-Fu Huang and Jhao-Min Hsu, "*Mining Web Logs to Improve Hit Ratios of Prefetching and Caching*", The 2005 IEEE International Conference on Web Intelligence (WI'05), Compiegne, France, pp. **577-580, 2005**.

[14] R.Sandrilla, M. Savitha Devi, "*A Study on Data Preprocessing Methods on Web Log Data in Web Usage Mining*", International Journal of Computer Sciences and Engineering, Vol.**6**, Issue.**7**, pp.**920-928, 2018.**

[15] AshirrK Kashyap, Iflah Naseem and Dheeraj Mandloi, "*Web Mining an Approach to Evaluate the Web*", International Journal of Scientific Research in Computer Science and Engineering, Vol.**5**, Issue.**3**, pp.**79-85**, **2017**.

[16] Sonia Sharma and Munishwar Rai, "*Customer Behaviour Analysis using Web Usage Mining*", International Journal of Scientific Research in Computer Science and Engineering, Vol.**5**, Issue.**6**, pp.47-50, 2017.

[17] Namrata Ghuse, Pranali Pawar and Amol Potgantwar, *"An Improved Approch For Fraud Detection In Health Insurance Using Data Mining Techniques"*, International Journal of Scientific Research in Network Security and Communication, Vol.**5**, Issue.**5**, pp.27-32, **2017**.

[18] M. Karthikeyan and P. Aruna, "*Probability based Document Clustering and Image Clustering using Content-based Image Retrieval*", Applied Soft Computing, Elsevier, Vol.**13**, Issue.**2**, pp.959-966, **2013.**

Authors Profile

Mr. K. Vadivazhagan is presently working as an Assistant Professor of Computer and Information Science and currently pursuing Ph.D. in Computer Science, Division of Computer and Information Science, Annamalai University, Annamalainagar, India. His research area focuses on Web Log Mining and Big Data Analytics. He has 16 years of teaching experience and 3 years of research experience apart from web developing.

Dr. M. Karthikeyan pursed Ph.D. in Computer Science and Engineering and presently working as an Assistant Professor in Division of Computer and Information Science, Annamalai University, Annamalainagar, India. He has published more than 15 research papers in reputed international Journals. His research area focuses on Data Mining, Digital Image Processing and Artificial Neural Networks. He has 19 years of teaching experience and 10 years of research experience.

