Data Preprocessing: A Crucial Pace for Pattern Discovery

Dr. Sonia Sharma

Associate Professor of C.Sc, Hindu Girls College, Jagadhri, Haryana, India

soniasharma1980@gmail.com

ABSTRACT: With the supersonic growth of internet and online business, many visitors visit the site of the organization of their choice from all over the world. In the current age of technology which is generally called e-business or e-commerce, to confront stiff competitiveness, and to explore consumer behavior is a puzzling task for the administrator of the organization. So to handle this puzzling task and to achieve objective of the organization web mining is used by most of the e-business organization. Mammoth records is stored on the web server in the form of weblogs on daily basis but this data does not clearly depict the information of the consumer visiting the website because it includes perplexing entries and it is herculean task to analyze and utilize this data for decision making. To explore jumbled and raucous data, web usage mining is a fruitful technique for sighting and for the scrutiny of web logs. To explore web usage mining in better way three prominent methods such as data preprocessing, pattern discovery and pattern analysis are used step by step. By applying these steps in proper way, a user can get the consumer browsing behavior of website. Because the exact information of user accessing the website is not clearly reflected and understood, therefore, web logs preprocessing is the crucial pace for pattern discovery. This paper designates the value of web-mining, process of web usage mining and data preprocessing method in an effective mode.

Key words: Web Mining, Data Pre-processing, Web Usage mining

I. INTRODUCTION

In this era of e-commerce knowing about the changing patterns of the preferences of the consumer is the most significant &challenging thing. In order to withstand cut-throat competitions Web mining is, indubitably, the most befitting method not only to withstand the cut through competitions but also to garner the dividends, which is the main motto behind any online business. Implementation of data mining applications i.e web mining [3] and web activities not only leads to the revelation of hidden information and extraction of desirable designs but also to the reduction in competitiveness and gaining the profits [4]. This paper emphatically highlights web usage mining (WUM) &Pre-processing phase of web usage mining because in this era to know the mind blowing patterns of consumers behaviour is an arduous task because majority of traders and industrialists firmly believe web is a baffling system only meant for the transactions .But unknowingly ,they fail to understand the worth of the data which comes into existence only after the interaction of the numerous visitors from all directions .The assembled data(weblogs) assists in decoding the puzzling behaviour of the consumers. Not only will i t bring improvement in consumer services and relationships and also in launching of the targets, forming marketing strategies and ultimately attainment of the goals [9][16]. Therefore investigating client's conduct is a substantial portion in the architecture of web .Detail of user sessions, browser, and operating system version [14] etc. is very important to know consumer behaviour and analysis of market [1] for the benefit of E-business. The information stored on the weblogs do not represent the clear picture of consumer interest [13]. So preprocessing of data is the most crucial step to reach out to the needs of the consumer. It is the procedure to transform the unorganized data stored within weblogs database for changing it into the format for further process of pattern discovery. Paper section- II describes the taxonomy of web mining. Step by step process of web usage mining is described in section –III labels the process of web usage mining. Data Pre-processing results are shown in section-IV and Section –V highlights the applications of web usage mining.

II. CLASSIFICATION OF WEB MINING

When the stored data is processed scientifically by using web mining the outcome is not interesting but also result producing. It can be defined as combination of World Wide Web and data mining. It consist of Data on the Web (content), Web log data (usage) and Web structure .Figure-I shows the classification of web mining.

 DOI : 10.35291/2454-9150.2023.0185

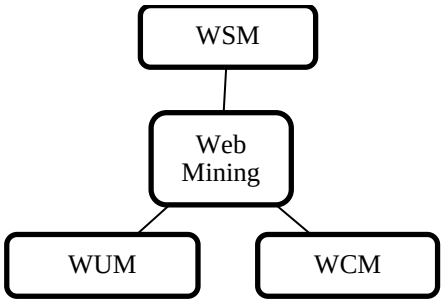


Figure-I: Web Mining Classification

a. **Web Content Mining(WCM)**

In web-data, on the basis of structured and unstructured data, semi structured data is reconstructed which is processed through web-content mining. WCM is the field to explore substantiation properties which accessed online.WCM is an established mean to get appropriate and authentic solutions and it includes preprocessing of search engine. Mining of text and multimedia files are included in WCM.

b. Web Usage Mining(WUM)

Web Usage Mining is extremely useful way for the in-depth study of a consumer's mind[15]. This mining unfolds within no time all its browsing information and all the trade related activities in a log file.

From server log file which leads a better understanding of consumers behaviour .like any other man made system ,the erection of the web can also be altered ,modified and upgraded depending upon users requirement [10]. During the process of consumer visits on the web important information is automatically received from the secondary data which is easily available on the log file by the different web usage methods. The goal of the mining which is foretelling consumer's behaviour within a website is strongly stresses by the learned author's. Normally consumer's general access patterns and individual usage records occur during web usage mining process. To improve the structure of website general access patterns are used and to explore the behaviour individual consumer's usage are used [5].

c. Web Structure Mining(WSM)

To procure all the relevant information regarding link available on the web page WSM is used. Because in any trade link information is crucial. It is generally observed textual information is used by most of the rescue tools but side line the bond material. After gauging the link of web page stayed by the consumers, it generates complete prototype of links available on web [11].

III. PROCESS OF WUM

Procedure of WUM flinch from collection of raw data, preprocessing of raw data, discovery of patterns and analysis of patterns. In web usage mining raw data (weblogs) which consist of consumer information is the main key to reach up to the consumer. Web site designer uses web server logs to know their consumer behaviour [11]. Stages of web usage mining is shown in figure-I

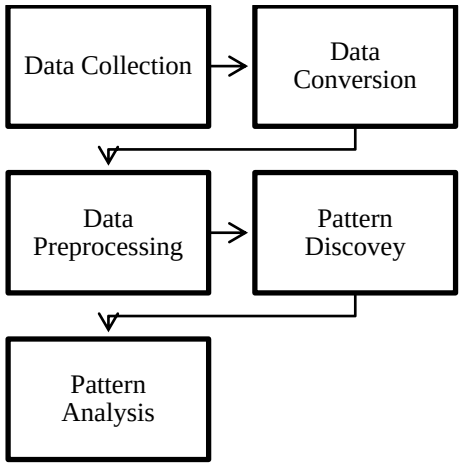


Figure-II: Stages of Web Usage Mining

a. Data Collection

To reach up to the consumer the first and most important step is collection of data. Web server, client server & internet cookies are three main sources for data collection. The main source of data collection is web server. Web server data is the main data to explore the consumer in an easiest way

b. Data Conversion

After collecting data from the web server, the next step is to transform the raw data (weblogs) into the layout which is easily understood through the administrator of the organization.

c. Data Preprocessing

To expand mining precision and to improve the quality of information, it is an essential step of WUM. It comprises cleaning of data and identification of user's sessions. Data pre-processing also known as information arrangement [12]. Steps of data preprocessing is shown in Figure-I.

Cleaning of data includes the amputation of irrelevant entries present within the data and sessions are identified by setting a threshold value. Web usage analytics can be done at this step to provide the answer of many queries such as which operating system is used? What is the most popular search engine? What request for the pages? And time spends by the consumers on the website.

 DOI : 10.35291/2454-9150.2023.0185

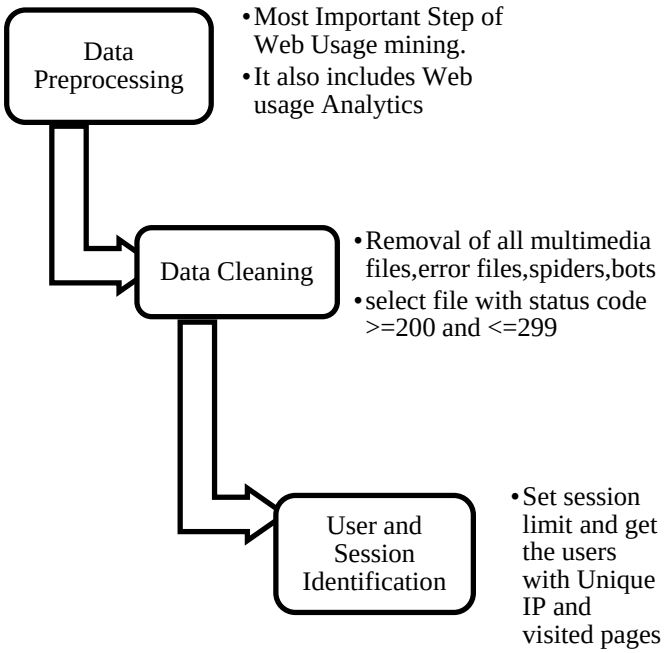


Figure-III: Steps of Data Preprocessing

d. Pattern Discovery

After session identification, next step under web usage mining is pattern discovery which is the most important step to know the relationship among data. It includes various techniques such as classification, clustering, association rule mining, Data analyst can apply various techniques of pattern discovery as per their need. The most popular technique for pattern discovery are association rule mining and clustering. For decision making and to help vendors in their marketing decisions. Partition clustering and hierarchical clustering methods of clustering technique can be used. To predict the consumer behaviour by knowing the most frequently visited pages, association rule mining is used. After knowing mining information web site design can be improved.[7].

e. Pattern Analysis

This is the last &imperative phase of web usage mining procedure. Clear analysis of results is very beneficial for an organization. Association can practice diverse policies such as online analytical programming (OLAP), Conception methods (Visualization),DKQ, and UA (Usability analysis) to analyses the patterns in an efficient manner. In this step results obtained during discovery step are analyzed and on the basis of which administrator can predict the behaviour of consumers and can check the objectives of the organization [11].

IV. RESULTS& DISSCUSSION

After indulgent the concept of web usage mining, the first step data preprocessing method is applied on real data (weblogs) of web site www.viralsach.xyz. Firstly the raw data is converted in .csv format then data cleaning and session identification algorithms are applied. Work is carried out in python version 3.2. Irrelevant entries are removed during data cleaning. Unique IP address are found during the session identification process. Work is done using Python Programming. For web usage analytics, weblog explorer lite 9.2 is used with which statistical analysis can be done such as popular search engine, browser version ,request method ,version of operating system etc. many results obtained during data preprocessing. Few of them are shown in figures IV, V, VI and VII.

Figure-IV depicts the number of rows (data size) obtained after data preprocessing. Data cleaning process is applied on 33172 log files and after data cleaning process 22129 rows are found and 18216 rows are obtained after session identification algorithms.

Figure-V depicts the usage of memory while performing data preprocessing and it is reduced from 1400Kb to 782.2Kb. Up to 50% memory is reduced during data preprocessing step.

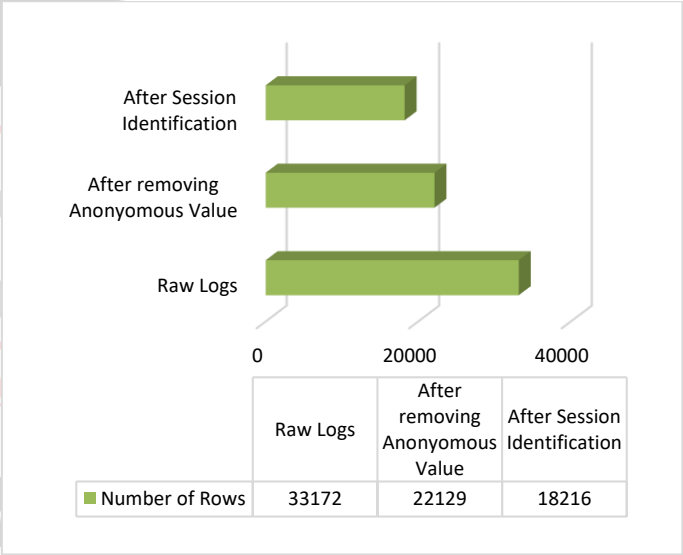
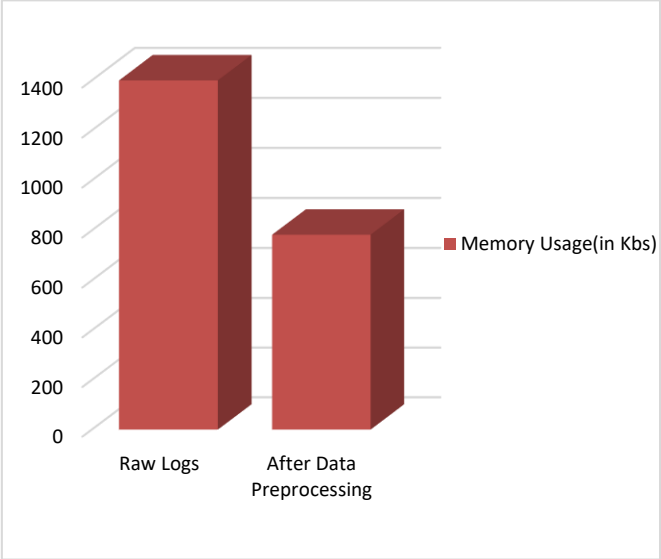


Figure-IV: Rows obtained after Data Pre-Processing



 DOI : 10.35291/2454-9150.2023.0185

Figure-VI depicts the most popular browser used by the visitors. It shows that 42.68% uses Google chrome and 26.68% Firefox and & Internet explorer is 18.58% used by the visitors and Microsoft edge is lease used by the visitors. In this way system administrator can predict from which browser consumers hits the site.

Figure-V: Difference in memory usage

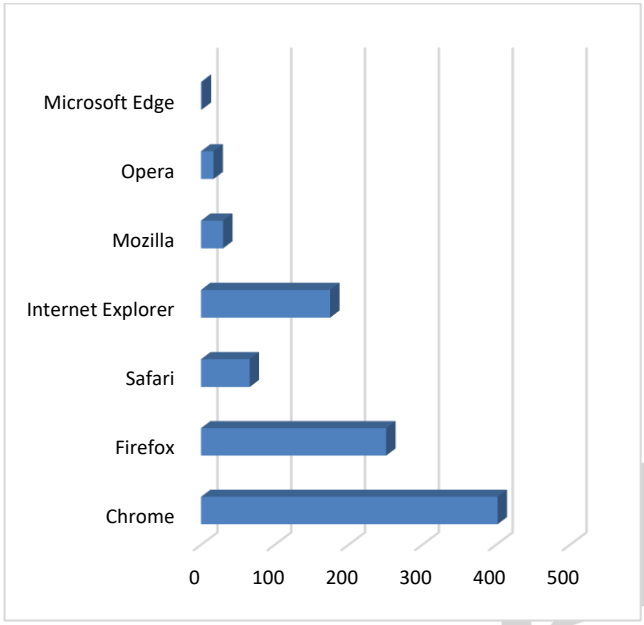


Figure-VI: Most used Browser

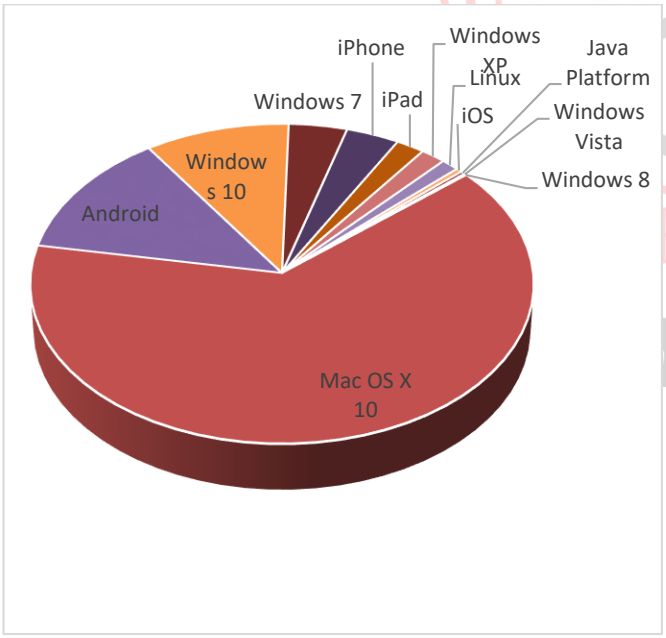


Figure-VII: Most Used Operating System

V. APPLICATIONS OF WEB USAGEMINING

After data preprocessing results, analyst can use pattern analysis and can take the benefits of this data for various application mentioned below because in the web based business commercial center, any extricated data of shopper conduct is significant to organizations.

Consumer Analyzing: Attracting new customers and side by side it helps retaining the old customers on the basis of processed and analyzed data. Trade administrator can increase their profit and achieve maximum benefit after analyzing the processed data [8].

Website enhancement: Optimized use of website not only facilitate the customers to provide clear vision of their habits, likes and characteristics but also the organizations to plan their strategies on the basis of those inputs.

Web Personalization: The procured data assist organization in multiple ways. On the basis of clear cut preferences as expressed by the customer the website can be redesigned and restructure infinitely .A trader can increase its administration worth and execution to fulfill its consumers by utilizing the conduct data [10].

Business Intelligent: In this competitive environment business can make better image by utilizing web usage mining and have heaps of potential for internet business promoting. Similarly objective of traders, industrialist can be achieved and can make business astute.

VI. CONCLUSION

In an online scenario web usage mining is most popular field used by e-business organizations to explore consumer behaviour to increase the profit and reach up to the consumer for business intelligence, decision making, website personalization, online learning. Full benefits of web usage mining can only be achieved by applying data preprocessing method efficiently and effectually. This stage is very crucial to know the behaviour of consumer. If this stage is not done in proper way then the discovery of pattern cannot be obtained in proper way and it will automatically effect the business of any organization. This paper describes the concept of data preprocessing by cleaning of data and identifications of sessions and web usage analytics. By knowing various results of memory consumption, anonymous values in weblogs, access statistics and usage analysis an administrator can take decision for changes in website as per the interest of consumer in using browser, search engine etc and a web site owner can easily decide the strategy to predict consumer behavior. After the completion of data preprocessing stage further pattern discovery techniques can be applied for the benefit of an organization

VII. REFERENCES

[1] Li Yong-hong; Liu Xiao-liang(2010), "Research of data mining based on e- commerce," Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on , vol.4, no., pp.719-722.

[2] Chaoyang Xiang; Shenghui He; Lei Chen (2009), International Symposium on A Studying System Based on Web Mining," Intelligent Ubiquitous Computing and Education, 2009, vol., no., pp.433-435.

 DOI : 10.35291/2454-9150.2023.0185

[3] Li Mei; Feng Cheng(2010), 2nd International Conference on Computer Engineering and Technology (ICCET), "Overview of Web mining technology and its application in e-commerce,", vol.7, no., pp.V7- 277,V7-280.

[4] Yadav, M.P.; Feeroz, M.; Yadav, V.K., "Mining the customer behavior using web usage mining in e-commerce," Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on , vol., no., pp.1,5, 26-28 July 2012.

[5] Zhiwu Liu; Li Wang, 3rd International Conference on Intelligent Networks and Intelligent Systems (ICINIS), 2010 "Study of Data Mining Technology Used for E- commerce,", vol., no., pp.509,512, 1-3 Nov. 2010

[6] ShenZihao; Wang Hui, International Conference on Intelligent Computing and Cognitive Informatics (ICICCI), 2010 "Research on E-Commerce Application Based on Web Mining,", vol., no., pp.337-340.

[7] Tang, Hewen; Yan, Honglin; Zengfang Yang; Yu Ma; Chunping Li, WASE International Conference on Information Engineering, 2009. ICIE '09."Application of Data Mining in Electronic Commerce", vol., no., pp.631, 634.

[8] http://www.pms.informatik.unimuenchen.de/lehre/projekt-diplom-arbeit/navigation-ack/doc/thesis.shtml.

[9] Spiliopoulou, M. (1999) "Data mining for the Web. In Proceedings of Principles of Data Miningand Knowledge Discovery", Third European conference, PKDD'99,P588-589

[10] Wang, y.(2000)"Web Mining and Knowledge Patterns",http://db.uwaterloo.ca/~tozsu/courses/cs748t/surveys/wang.pdf.

[11] Cooley, Bamshad and Jaideep (1997), "Web Mining: Information and Pattern Discovery on thewwwhttp://wwwusers.cs.umn.edu/~mobasher/webminer/survey/survey.htm l.

[12] Jozef K, Michal M, Martin D, "User Session Identification Using Reference Length" in 9th International Scientific Conference On Distance Learning In Applied Informatics;2012 pp-175-184.

[13] Morzy T, Wojcie M, and Zakrazewicz M. "Web Use Clustering" International Symposium On Computer and Information Sciences, 2000

[14] Catlegde L. and Pitkow J., "Characterizing browsing behaviors in the world wide Web," Computer Networks and ISDN systems, 1995.

[15] Sonia Sharma, Dalip(2020), "A Novel Secure Web Usage Mining Technique to Predict Consumer Behaviour" International Journal of Advanced Science and Technology. Vol. 29, No. 5, (2020), pp. 5633 – 5640.ISSN: 2005-4238 IJAST.

[16] Sonia Sharma, Dalip(2019),"Comparative Analysis of various tools to Predict Consumer Behaviour" Journal of Computational and Theoretical Nano science Vol. 16, 3860–3866, 2019

 DOI : 10.35291/2454-9150.2023.0185