An Overview of Web Usage Mining: Methods, Uses, and Difficulties

Vinay shah Uikey#1, Arun Kumar Rai2

1Department of Computer Science and Engineering, Vedica Institute of Technology, Bhopal India,

2Department of Computer Science and Engineering, Vedica Institute of Technology, Bhopal India,

1vinayshah6063@gmail.com

2raiaruniitr@gmail.com

Abstract*— Web mining is the application of data mining methods to extract meaning from online resources, such as web pages, link structures, or the logs of user interactions. The three primary areas of web usage mining—web contents mining, web design mining operations, and web usage mining—are the focus of this review paper's investigation of methods and applications. Web usage mining specifically seeks to examine user behaviour patterns in web logs in order to enhance online designs, customize user experiences, and maximize web server performance. This work tackles the intricacies of data pre-processing, which is another challenge.

Keywords — Web log analysis, user behaviour analysis, data mining, and web mining.

I. Introduction

Web mining uses data mining techniques to extract valuable information from online data. Web pages, their links, and user interaction records on websites are all included in this. Web mining finds important trends and patterns by examining large online datasets. The details, papers, structures, and user profiles are the different categories into which this extracted data can be separated. Web mining is based on both data-driven and process-based components. Data collection, selection before processing, pattern recognition, and analysis are some of the phases it includes. Because of the internet's widespread use in modern life, research on methods for retrieving information from the internet has grown significantly. These techniques are especially effective when they are used to mine structure or usage data (such as Weblogs). As seen in figure 1, data mining activities can be roughly divided into three types. While each category focuses on distinct facets of web data, they all seek to uncover hidden, important information. For instance, web content mining explores web page content to extract information, details, and insights. Text, images, and groups of websites are all examined and analysed, and the results are then generated for search engines. On the other hand, web structure mining struggles with the complexities of web linking structures. Despite being a traditional field, link analysis has seen a resurgence in interest due to the rise of web mining.

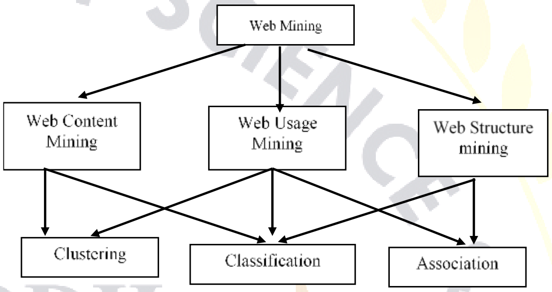

Figure 1 Web Data Mining Taxonomy

II. Mining Web Usage

The methods used in this domain are focused on predicting how users would behave when using the internet. Web use mining is used to identify trends in user navigation using online data. This type of mining looks for valuable insights in the additional data that is received while people browse the internet. The present form of mining allows one to identify trends in the way consumers visit specific websites by extracting data from web logs. Both commercial products and current research endeavors seek pattern that can be analyzed for a variety of purposes. Website personalization, system upgrades, website modifications, company analysis, and even customer classification can all benefit from these analytical findings. online usage mining, often known as log mining, is the process of tracking user activity on websites and gathering this data from online logs [1]– [3].

The internet protocol (IP) address, visit time stamp, and pages seen are among the data that is saved following a user's visit to a website. These specifics are gathered, examined, and recorded in logs. [4] [5]. Because it makes it easier to examine user activity, it is used to make sure that users' experiences on interactive websites are improved. [6]. This tool uses technology to automatically track users' access patterns. Web servers provide access to a large portion of this data, which is subsequently summarized in access logs. URLs, visiting times, internet protocol addresses, and additional data that could assist a business better understand its customers' behaviour and provide excellent customer care are all included in logs. [7] [8] Data from logs, such as user access patterns, are collected and analyzed via web usage mining. Web use mining's primary goal is to monitor the actions of users as they engage with the web. There are two approaches to pattern analysis: generalised tracking and customized tracking. Tracking information is frequently gathered from web page history. [9]

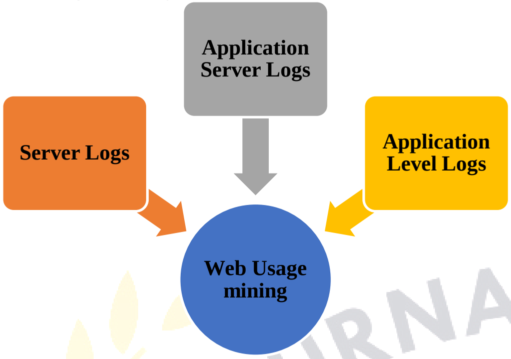

Figure 2 Data source for web usage mining

Internet usage mining is the process of using data mining techniques to uncover significant usage patterns in web data in order to better understand and meet the demands of web applications [10–11]. It is commonly referred to as the process of creating user behaviour patterns using the log files along with other pertinent information processed on particular websites.

Pre-processing, finding and assessing patterns, and assessing these patterns are the three stages of web usage mining. To gain a deeper understanding of visitor behaviour, methods such path assessment, association rules, clustering, classification, and sequential examination are employed [12–13].

Web usage data primarily serves to document the typical browsing patterns of a website's users. Web usage mining can be categorized based on the type of usage data examined [14-15]. In our study, we focus on web log data that chronicles user navigation patterns, as depicted in figure 1.2. Centered around web usage mining, our approach utilizes data mining techniques to uncover trends in web usage data. Users' interactions with the web result in the generation of data mainly in the form of web/proxy server logs. Usage mining methods attempt to study and predict user's behavior, so designers of web sites may make improvements into user's experience while offering an attractive web site, attract more visitors, and provide personalized services to devoted users. However, the nature of the data forming the basis of Web Usage Mining therefore presents a very real challenge. The growth of the internet meant exponential quantities of web data and most transactions take place in milliseconds. This vast amount of data is often semi-structured, making it distinct from traditional structured data. The growth of the internet meant exponential quantities of web data and most transactions take place in milliseconds. This vast amount of data is often semi-structured, making it distinct from traditional structured data [16]. Therefore, it demands extensive preprocessing to extract meaningful information from this semi-structured content.

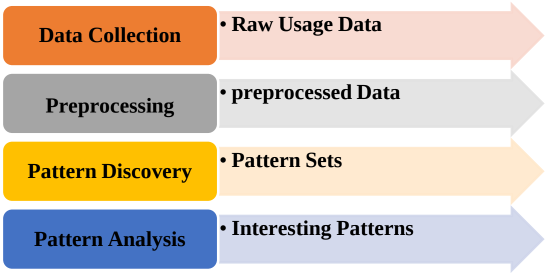

Figure 3 Web Usage Mining Phases

Weblog files from web servers are the main source of data for web usage mining, as seen in Figure 3. Web servers, Web proxy servers, and client browsers are the three primary sources from which this can be acquired. Although web servers can be set up to log a variety of information, the majority do so for IP addresses, login credentials, last page viewed, success rate, user agents, and URLs [17–18]. But the data from these logs is frequently vague and lacking.

About 80% of the processing time is spent on pre-processing, a crucial step in online usage mining that cleans, validates, and completes the data [19–20]. The process of turning pre-processed data into useful patterns is intricate and usually involves methods like regression analysis, clustering, and classification. The last step, pattern analysis, focuses on identifying and keeping intriguing patterns while eliminating those that are not important s[21].

III. Literature Review

Gholizadeh et al. (2021) [22] used association rules to analyze a relationship between production rules in the manufacturing lines. They addressed the logical reasoning interaction between the automobile production lines and goods with this method. The results demonstrated that an accuracy rate exceeding 87% was achieved with the application of data mining association rules. The method is helpful for simplifying the production process, enhancing management decision-making in the IoT domain, and producing manufacturing guidelines.

Asadianfam et al. (2020) [23] suggested webpage suggestions with c-means clustering separated a random user sample and carried out clustering using the information gathered. When compared to alternative approaches, the experiment's results utilizing this model produced a list of the user's top recommended pages with the highest accuracy of 89.56% and the highest precision of 96.45%.

Mohanty et al. (2023) [24] grouped the fuzzy membership matrices using the traditional fuzzy clustering algorithm, fuzzy c-means, and presented the clustering process' results. A new equation for the expression of fuzzy membership vectors is used in conjunction with the feature weight calculation method used in the clustering phase. A 10% increase in clustering accuracy is shown when comparing the experimental findings to the previous clustering techniques used on the multivariate time series datasets.

H. Li and M. Wei, (2024) [25] centered on creating algorithms for continuous mining, specifically for density-based clustering. These techniques are reliable for finding clusters with different granularities or eliminating outliers from areas with variable densities. On specific datasets, their accuracy rate was 60.375%.

P. Bhattacharjee et al. (2022) [26] highlighted artificial intelligence research, showing how conventional clustering techniques can be used with creative approaches to gain increased efficiency. They concentrated on clustering methods like K-means, CLARA, and CURE. The effectiveness of their approach was substantially greater than that of its raw data output. Additionally, they included a comparison with results from other approaches that used the same dataset, based on the Rand index. Interestingly, the CLARA approach outperformed the DBSCAN method on one dataset and reached 100% efficiency on two others.

M. Kotyrba et al. (2021) [27] suggested CLARANS-based clustering techniques. The unique feature of their method is that it only takes into account users who behave consistently throughout clustering. The quality and speed of clustering have significantly improved, according to the results. A silhouette coefficient of 0.7830 and notable decreases in different computing requirements are among the metrics.

M. S. Bhuvaneswari and K. Muneeswaran (2020) [28] identified patterns in inspection reports by text mining based on rules. They used unsupervised clustering to categorize projected renewal occurrences into four categories of criticality based on collected data and other important factors, such age. Their methodology indicates that just 8.8% of damaged roofs are of the highest importance and fall within the allocated budget. This flexible approach can be used for a variety of assets and can greatly improve the way major owner organizations, such as school boards and municipalities, allocate funds.

B. Gao et al. (2022) [29] presented a method for finding web blogs that uses an adaptive fuzzy feedback recurrent neural network. Using a fuzzy-based RNN, the method does rerank the search results, and user feedback helps it get even closer to what the user intended. Metrics like accuracy, precision, and recall were used to assess the efficacy of the Python implementation. With an accuracy rate of 94%, it outperforms well-known methods like ANN, DNN, and deep auto-encoders.

Khatter et al. (2020) [30] suggested using the least square regression technique to find patterns in sensor data analysis. To increase modeling efficacy and resilience in categorization, they refined the least squares regression technique, enabling test accuracy rates of roughly 96.2%s and 100%. These findings imply the effectiveness of the suggested sensor system.

Applications of Web Usage Mining

Discovering patterns in browsing behavior is one of the main objectives of web usage mining, and by doing so, websites can become more appealing to users. Three primary uses are as follows:

Privatization of Web Content: - User experience personalization will change as a result of Web usage mining techniques. It may be possible to utilize predictive algorithms to infer in real time what a user could do next by comparing the user's current browsing activity with log data from the past.

Thus, this guarantees that recommendation systems are able to suggest material and direct users to websites that are most appropriate for their interests. This makes it possible to better facilitate the improvement of the user experience by customizing the product to be offered and organizing catalogs based on the user's most likely interests. In addition to improving user pleasure, this customisation raises user engagement and interaction with the web content.

Pre-recovery: - Web usage mining can offer vast performance enhancement to web servers and web-based applications. Techniques developed using data mining techniques can be used for developing efficient retrieval and caching techniques by the servers. Optimization of retrieval times reduces response time for web requests and thereby improves user experience as a whole. Improved efficiency in using a server improves responsiveness and reliability of web applications through optimum pré-emptive caching of frequently accessed content and dynamic server response to usage patterns. Its proactive approach toward server management ensures smoother interaction and fast loading times for users accessing web services.

Optimization of Web Design: - The usability parameter is very crucial in developing the effective web sites. The web usage mining provides excellent information that can be used to enhance the design and structure of a web page. For example, the collected data regarding user interaction can be put to use through the development of an adaptive web page, where the pages change with respect to the observed behaviors. In this dynamic approach, website content will relate and interest end-users over time. Usage analytics would assist in improving design elements along with the form of presentation of content on website. In case the said usage analytics are being constantly used, it will ensure that engagement and satisfaction by the users remain at the optimal level. Thus, organizations would be able to enhance usability and navigation and improve their overall user experiences from websites, driving a higher level of user retention and interaction.

IV. Conclusions

Web usage mining helps uncover precious insights from web log data so that true user behavior understanding may be achieved. Different techniques such as clustering, classification, and association rules enable organizations to deliver content better, improve their web server's efficiency, and fine-tune their web design according to the interactions between the user and the web site. However, along with the advantages it brings regarding utilization of web usage mining techniques, it cannot sideline the hurdles including processing semi-structured data and too much preprocessing. Therefore, future studies should be conducted to overcome these obstacles and increase the usability of web usage mining in web analytics and related user-centered applications.

REFERENCES

[1] P. Verma and N. Kesswani, “FEDUS: A comprehensive algorithm for web usage mining,” https://doi.org/10.1080/02522667.2019.1616912, vol. 41, no. 3, pp. 835–854, Apr. 2019, doi: 10.1080/02522667.2019.1616912.

[2] B. Sheykh Abbasi, N. Abdolvand, and S. Rajaee Harandi, “Predicting Customers’ Behavior Using Web-Content Mining and Web-Usage Mining,” Int. J. Inf. Sci. Manag., vol. 20, no. 3, pp. 141–163, Jul. 2022, Accessed: May 03, 2023. [Online]. Available: https://ijism.ricest.ac.ir/article_698405.html.

[3] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. Niakan Kalhori, “Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study,” JMIR Public Heal. Surveill., vol. 6, no. 2, Apr. 2020, doi: 10.2196/18828.

[4] M. Babagoli, M. P. Aghababa, and V. Solouk, “Heuristic nonlinear regression strategy for detecting phishing websites,” Soft Comput., vol. 23, no. 12, pp. 4315–4327, Jun. 2019, doi: 10.1007/S00500-018-3084-2/METRICS.

[5] C. Giri, S. Thomassey, J. Balkow, and X. Zeng, “Forecasting New Apparel Sales Using Deep Learning and Nonlinear Neural Network Regression,” 2019 Int. Conf. Eng. Sci. Ind. Appl. ICESI 2019, Aug. 2019, doi: 10.1109/ICESI.2019.8863024.

[6] N. Deepika and M. N. Bhat, “Predicting the E-Commerce Companies Stock with the Aid of Web Advertising via Search Engine and Social Media,” Rev. d’Intelligence Artif., vol. 34, no. 1, pp. 89–94, 2020, doi: 10.18280/RIA.340112.

[7] X. Li and Z. Li, “A hybrid prediction model for e-commerce customer churn based on logistic regression and extreme gradient boosting algorithm,” Ing. des Syst. d’Information, vol. 24, no. 5, pp. 525–530, 2019, doi: 10.18280/ISI.240510.

[8] Fei Li, Yun Li, "Big Data Mining Method of E-Commerce Consumption Pattern Based on Mobile Platform", Security and Communication Networks, vol. 2022, Article ID 3991135, 12 pages, 2022. https://doi.org/10.1155/2022/3991135

[9] F. Xu and S. Qu, “Data Mining of Students&rsquo; Consumption Behaviour Pattern Based on Self-Attention Graph Neural Network,” Appl. Sci. 2021, Vol. 11, Page 10784, vol. 11, no. 22, p. 10784, Nov. 2021, doi: 10.3390/APP112210784.

[10] O. I. Abiodun et al., "Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition," in IEEE Access, vol. 7, pp. 158820-158846, 2019, doi: 10.1109/ACCESS.2019.2945545

[11] H. K. Sowmya and R. J. Anandhi, “An efficient and scalable dynamic session identification framework for web usage mining,” Int. J. Inf. Technol., vol. 14, no. 3, pp. 1515–1523, May 2022, doi: 10.1007/S41870-022-00867-3.

[12] Suragala, Ashok, P. Venkateswarlu, and M. China Raju. "A comparative study of performance metrics of data mining algorithms on medical data." ICCCE 2020: Proceedings of the 3rd International Conference on Communications and Cyber Physical Engineering. Springer Singapore, 2021.

[13] S. Umamaheswari and K. Harikumar, “Analyzing product usage based on twitter users based on datamining process,” Proc. Int. Conf. Comput. Autom. Knowl. Manag. ICCAKM 2020, pp. 426–430, Jan. 2020, doi: 10.1109/ICCAKM46823.2020.9051488.

[14] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, “Classification of tweets data based on polarity using improved RBF kernel of SVM,” Int. J. Inf. Technol., vol. 15, no. 2, pp. 965–980, Feb. 2020, doi: 10.1007/S41870-019-00409-4/METRICS.

[15] Karthikeyan, T., Karthik Sekaran, D. Ranjith, and J. M. Balajee. "Personalized content extraction and text classification using effective web scraping techniques." International Journal of Web Portals (IJWP) 11, no. 2 (2019): 41-52.

[16] H. K. Sowmya and R. J. Anandhi, “An efficient and scalable dynamic session identification framework for web usage mining,” Int. J. Inf. Technol., vol. 14, no. 3, pp. 1515–1523, May 2022, doi: 10.1007/S41870-022-00867-3.

[17] G. Chaubey, P. R. Gavhane, D. Bisen, and S. K. Arjaria, “Customer purchasing behavior prediction using machine learning classification techniques,” J. Ambient Intell. Humaniz. Comput., pp. 1–25, Apr. 2022, doi: 10.1007/S12652-022-03837-6/METRICS.

[18] C. Nigam and A. K. Sharma, “WITHDRAWN: Experimental performance analysis of web recommendation model in web usage mining using KNN page ranking classification approach,” Mater. Today Proc., Oct. 2020, doi: 10.1016/J.MATPR.2020.09.364.

[19] Yağcı, Mustafa. "Educational data mining: prediction of students' academic performance using machine learning algorithms." Smart Learning Environments 9, no. 1 (2022): 11.

[20] L. Wang, B. Lin, R. Chen, and K. H. Lu, “Using data mining methods to develop manufacturing production rule in IoT environment,” J. Supercomput., vol. 78, no. 3, pp. 4526–4549, Feb. 2022, doi: 10.1007/S11227-021-04034-6/METRICS.

[21] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, “Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest,” Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, pp. 679–684, Mar. 2019, doi: 10.1109/ICCMC.2019.8819654.

[22] Gholizadeh, Nahid, Hamid Saadatfar, and Nooshin Hanafi. "K-DBSCAN: An improved DBSCAN algorithm for big data." The Journal of Supercomputing 77 (2021): 6214-6235.

[23] S. Asadianfam, H. Kolivand, and S. Asadianfam, “A new approach for web usage mining using case based reasoning,” SN Appl. Sci., vol. 2, no. 7, pp. 1–11, Jul. 2020, doi: 10.1007/S42452-020-3046-Z/TABLES/3.

[24] S. N. Mohanty, J. Rejina Parvin, K. Vinoth Kumar, K. C. Ramya, S. Sheeba Rani, and S. K. Lakshmanaprabu, “Optimal rough fuzzy clustering for user profile ontology based web page recommendation analysis,” J. Intell. Fuzzy Syst., vol. 37, no. 1, pp. 205–216, Jan. 2019, doi: 10.3233/JIFS-179078.

[25] H. Li and M. Wei, “Fuzzy clustering based on feature weights for multivariate time series,” Knowledge-Based Syst., vol. 197, p. 105907, Jun. 2020, doi: 10.1016/J.KNOSYS.2020.105907.

[26] P. Bhattacharjee and P. Mitra, “Density-Based Mining Algorithms for Dynamic Data: An Incremental Approach,” Stud. Comput. Intell., vol. 1028, pp. 313–335, 2022, doi: 10.1007/978-981-19-1021-0_13/COVER.

[27] M. Kotyrba, E. Volna, R. Jarusek, and P. Smolka, “The use of conventional clustering methods combined with SOM to increase the efficiency,” Neural Comput. Appl., vol. 33, no. 23, pp. 16519–16531, Dec. 2021, doi: 10.1007/S00521-021-06251-9.

[28] M. S. Bhuvaneswari and K. Muneeswaran, “User Community Detection From Web Server Log Using Between User Similarity Metric,” Int. J. Comput. Intell. Syst., vol. 14, no. 1, pp. 266–281, 2020, doi: 10.2991/IJCIS.D.201126.002.

[29] B. Gao, “The Use of Machine Learning Combined with Data Mining Technology in Financial Risk Prevention,” Comput. Econ., vol. 59, no. 4, pp. 1385–1405, Apr. 2022, doi: 10.1007/S10614-021-10101-0/METRICS.

[30] H. Khatter and A. K. Ahlawat, “An intelligent personalized web blog searching technique using fuzzy-based feedback recurrent neural network,” Soft Comput., vol. 24, no. 12, pp. 9321–9333, Jun. 2020, doi: 10.1007/S00500-020-04891-Y/METRICS.