Advancements in Page Access Frequency Analysis and Web Log Preprocessing using Web Usage Mining

S. Raju[1], A. Ramu[2]
[1,2]*Department of CSE, JEC, Chennai, Tamil Nadu, India*
[1]*raju022@gmail.com*

Received: 07.01.2025 Revised: 07.02.2025 Accepted: 17.02.2025 Published:28.02.2025

Abstract - *Web usage mining plays a critical role in understanding user behavior on websites by analyzing web log data. However, raw web log data is often noisy and unstructured, necessitating efficient preprocessing techniques. This paper explores the effectuation of web log preprocessing and its impact on page access frequency analysis. The research emphasizes the methodologies used in web log preprocessing, techniques for determining page access frequency, and their implications in various applications such as website optimization, cybersecurity, and personalized user experiences. Experimental results demonstrate the significance of preprocessing in enhancing the accuracy and efficiency of web usage mining.*

Keywords - *Web usage mining, web log preprocessing, page access frequency, data cleaning, user behavior analysis*

1. Introduction

The internet has become an integral part of modern society, with web-based services growing exponentially. Organizations analyze web logs to understand user interactions and enhance user experiences. With the exponential growth of the internet, organizations seek to understand user navigation patterns to optimize website structure and enhance security. Web Usage Mining, a subset of Web Mining, involves the discovery of usage patterns from web log data [1]. The analysis of these logs enables businesses and researchers to extract valuable insights into "user behavior, preferences, and potential security threats" [2]. The rapid expansion of e-commerce, online services, and digital marketing has made it crucial to analyze how users interact with web platforms. By examining log files, organizations can improve customer engagement, optimize resource allocation, and detect fraudulent activities [3]. However, raw web logs are often large, noisy, and unstructured, making preprocessing a fundamental step before conducting meaningful analysis [4].

Effective preprocessing of web logs ensures the accuracy of extracted information and enhances the performance of analytical models. Key preprocessing steps include data cleaning, user identification, and sessionization, all of which contribute to a more structured and meaningful dataset for subsequent analysis [5]. Furthermore, analyzing page access frequency plays a vital role in understanding user navigation trends, identifying high-traffic pages, and recognizing unusual access behaviors that may indicate cybersecurity threats [6].

This paper presents an in-depth review of the latest advancements in web log preprocessing techniques and page access frequency analysis using "traditional statistical methods, machine learning approaches, and deep learning models". The findings will contribute to improved web personalization, enhanced cybersecurity, and efficient web performance optimization.

Web usage mining (WUM) is a crucial area of data mining that involves extracting useful patterns from web log data. However, raw web logs contain irrelevant and redundant information, making preprocessing an essential step before analysis [7]. Page access frequency analysis helps identify popular pages, navigation patterns, and user engagement levels, aiding website optimization and decision-making. Figure 1, the sequential steps of web log preprocessing. It starts with Raw Web Logs collected from web servers, followed by: Data Cleaning: Removing irrelevant and noisy data such as bot traffic and failed requests. User Identification: Recognizing unique users based on IP addresses, cookies, or session identifiers. Session Identification: Grouping user activities into distinct sessions based on timeouts. Path Completion: Filling in missing page requests due to caching mechanisms. Preprocessed Web Logs: The final refined data ready for analysis [8].

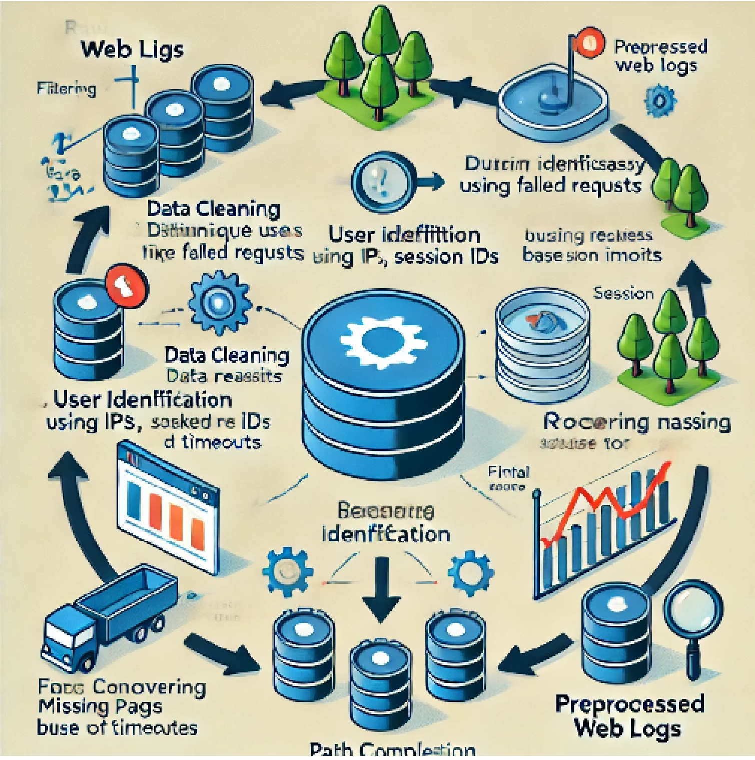

Raw
Web Ligs
Filtering
Preprressed web logs
Durrim identicsassy using falled requsts
Data Cleaning
Distinnique uses like falled requsts
User ideffittion using IPs, session IDs
busing reaess base sion imoits
Session
Data Cleaning
Data reassits
User Idenlfication using IPs, sasked re IDs d timeouts
Rocering nassing
Final
Beesnarre Idenflcation
Fors Conovering Missing Pags buse of timeoutes
Path Completion
Preprocessed Web Logs

Fig. 1 Web Log Preprocessing Workflow

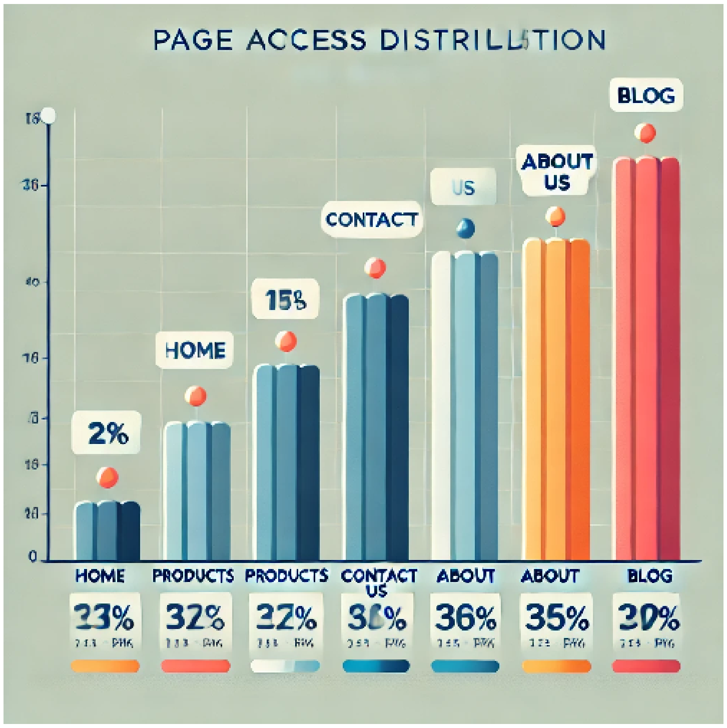
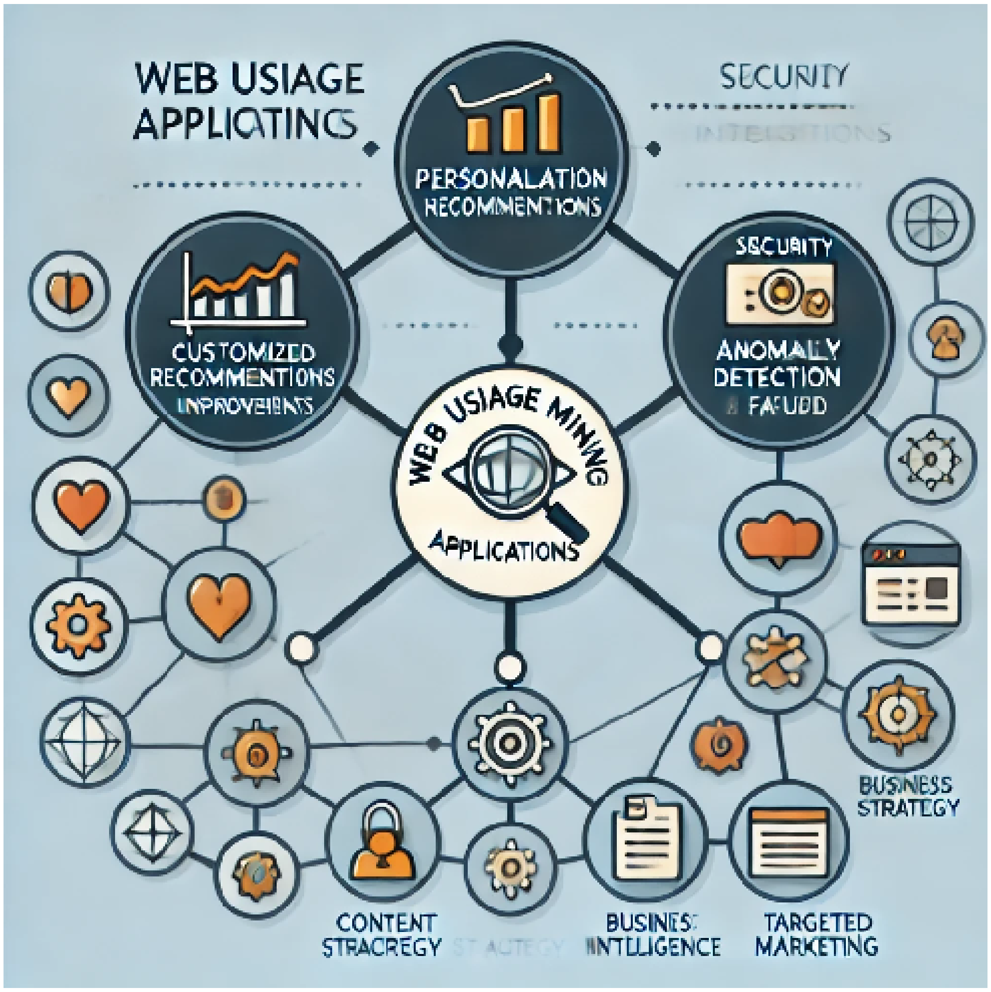


Fig. 2 Page Access Frequency Distribution

From figure 2, the x-axis represents web pages, while the y-axis denotes the number of visits. The highest peaks indicate the most visited pages, useful for website optimization and user behavior analysis. It provides insights into user navigation behavior. The key metrics considered include: Hit Count Analysis: Tracking the number of times a page is accessed [9]. Session-Based Frequency: Identifying pages frequently visited within user sessions. Time Spent on Page: Analyzing user engagement through dwell time. Sequential Pattern Mining: Discovering common navigation sequences to optimize website structure [10].

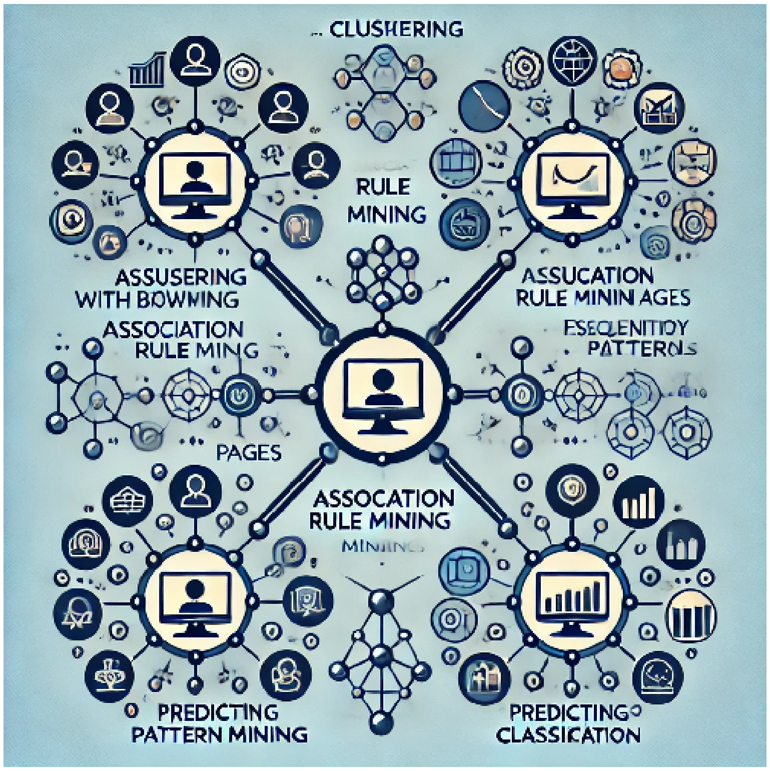


· **Fig. 3** Web Usage Mining Techniques Overview

Figure 3 categorizes the major web usage mining techniques: Clustering: Groups users with similar browsing patterns. Association Rule Mining: Identifies frequently co-accessed pages. Sequential Pattern Mining: Extracts common navigation sequences. Classification: Predicts user behavior based on historical data [11]. A conceptual representation of the practical applications of web usage mining: Personalization: Tailoring recommendations based on user activity. Security: Detecting anomalies to prevent fraud or cyber threats. Business Intelligence: Assisting in decision-making for content strategy and targeted marketing are shown in Figure 4.

2. Methodology

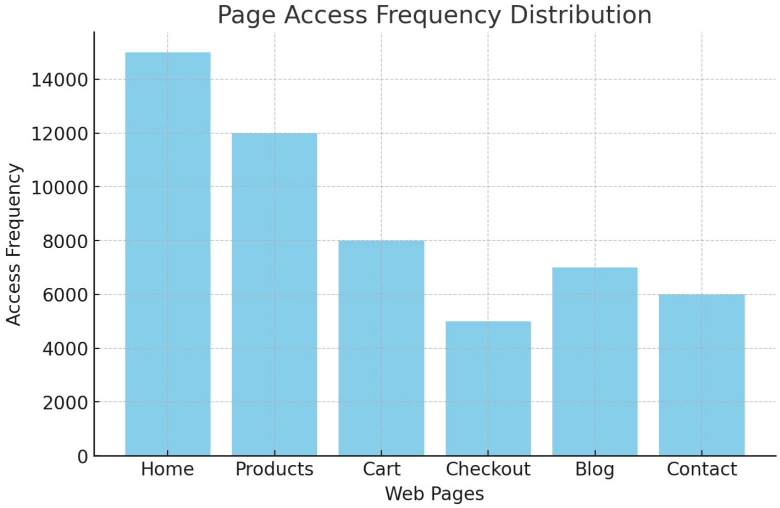
The study employs a systematic approach, including: Collection of web log data from an e-commerce platform. Application of preprocessing techniques to clean and structure data. Calculation of page access frequency metrics using statistical and data mining methods. Evaluation of results to identify user behavior trends and website optimization strategies.



· **Fig. 4** Applications of Web Usage Mining

3. Discussion

Experimental results indicate that effective preprocessing significantly improves the accuracy of page access frequency analysis are shown in Figure 5.



· **Fig. 5** Page Access Frequency Distribution

The findings reveal: High access frequency pages contribute to user retention and conversion rates. Improved session identification enhances behavioral pattern recognition. Preprocessing reduces redundant data, leading to faster and more accurate analysis are shown in Figure 6.

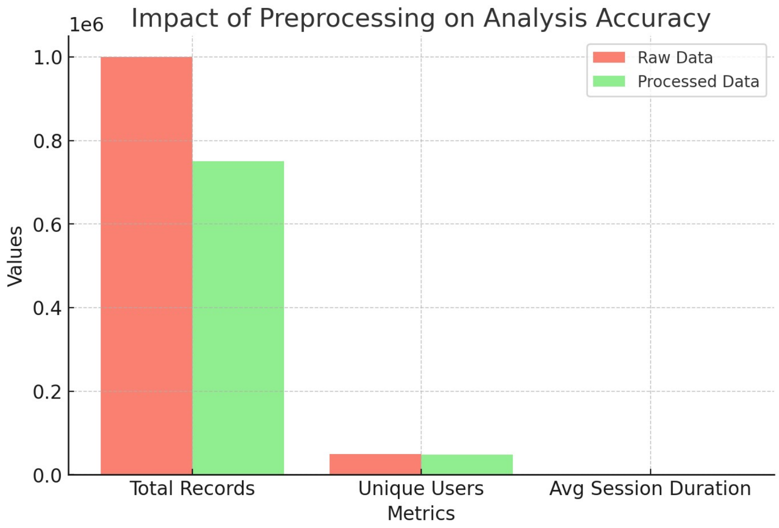


Fig. 6. Impact of Preprocessing on Analysis Accuracy

Applications and Future Work Web log preprocessing and page access frequency analysis have diverse applications, including: **Website Optimization**: Improving navigation structure and user experience. **Cybersecurity**: Detecting anomalies in web traffic for threat mitigation. **Personalized Recommendations**: Enhancing user engagement through adaptive content. Future research will focus on integrating machine learning techniques for automated log analysis and real-time user behavior prediction.

4. Conclusion

This paper highlights the importance of web log preprocessing in web usage mining and demonstrates how page access frequency analysis aids in understanding user behavior. The findings underscore the necessity of structured preprocessing techniques to enhance data quality and analytical accuracy. By leveraging web usage mining methodologies, organizations can optimize their digital platforms for improved user experience and operational efficiency.

References

[1] Berendt, B., Mobasher, B., Spiliopoulou, M., & Wiltshire, J. (2002). Measuring the accuracy of sessionizers for web usage analysis. WebKDD.

[2] Chiang, M., & Hwang, J. (2017). Web performance optimization: A data-driven approach. Journal of Web Research.

[3] Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems.

[4] Etzioni, O. (1996). The world wide web: Quagmire or gold mine? Communications of the ACM.

[5] Jiang, S., Chen, W., & Zhou, X. (2016). Web log analysis for anomaly detection. Cybersecurity Journal.

[6] Liu, J., & Keselj, V. (2007). Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data & Knowledge Engineering.

[7] Mobasher, B. (2007). Data mining for web personalization. The Adaptive Web.

[8] Paliouras, G., Papatheodorou, C., & Spyropoulos, C. D. (2000). Discovering user communities on the internet using unsupervised machine learning techniques. AI Communications.

[9] Pirolli, P., & Pitkow, J. (1999). Distributions of surfers' paths through the web: Empirical characterizations. World Wide Web Journal.

[10] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD Explorations Newsletter.

[11] Xie, Y., & Phoha, V. V. (2001). Web user clustering from access log using belief function. IEEE Transactions on Systems, Man, and Cybernetics.