10.48047/jocaaa.2024.33.08.242

Advanced Server-Side Data Collection Methodology for Web Usage Mining: Eliminating Preprocessing Overhead and Enhancing Analytics Accuracy

Mohit Paul[1], Gaurav Pandey[2], Abdul Rub[3]

[1,2,3] **Department of Computer Science and Information Technology, SHUATS, Prayagraj**
mohit90.paul@gmail.com[1*], gauravpandey161989@gmail.com[2], abdulrub88@gmail.com[3]

Abstract

This research introduces a novel server-side methodology for collecting web usage data that addresses the limitations of traditional preprocessing-heavy techniques. Unlike conventional approaches that rely primarily on web server access logs, the proposed system eliminates complex preprocessing while improving data quality, supporting real-time analytics, and ensuring stronger privacy controls. Standard server logs offer only fragmented client information, making session identification laborious and often inaccurate, which in turn hinders their suitability for effective web analytics.[1] To overcome these challenges, we present an application-driven API framework designed to track users, manage sessions, and capture usage data in a structured manner. Unlike client-side tools, this solution operates seamlessly within enterprise web applications to gather and process detailed interaction records. The homogeneity and organization of the collected data make it significantly easier to filter, analyze, and apply across various domains such as large-scale web usage mining, real-time analytical systems, intelligent recommendation engines, and machine learning pipelines [3]. Experimental results confirm that this method delivers high-performance, structured, and readily usable data compared to conventional log-based approaches.

Introduction

Web Usage Mining (WUM) has steadily transformed into a cornerstone of digital analytics, offering organizations the ability to unravel subtle patterns and hidden trends in user interactions across the web. Unlike conventional analytics, which often stops at surface-level metrics, WUM dives deep into the behavioral footprints left behind by users, thereby providing actionable intelligence for personalization, recommendation, and adaptive web design [1]. Historically, the foundation of WUM rested heavily on raw server access logs—massive, unstructured datasets that, while abundant, introduced formidable challenges. These challenges are most pronounced in the preprocessing stage, often described as the “heart and bottleneck” of WUM, consuming nearly 60–80% of the total mining effort. Here, tedious yet indispensable tasks such as noise reduction through data cleaning, accurate user identification across shared IPs, robust session tracking in the absence of explicit boundaries, reconstructing incomplete navigation paths, and formatting heterogeneous data sources must be meticulously executed. This preprocessing complexity not only delays the analytical process but also determines the reliability and precision of the final mined patterns, making it one of the most peculiar yet pivotal phases of the entire WUM pipeline.

10.48047/jocaaa.2024.33.08.242

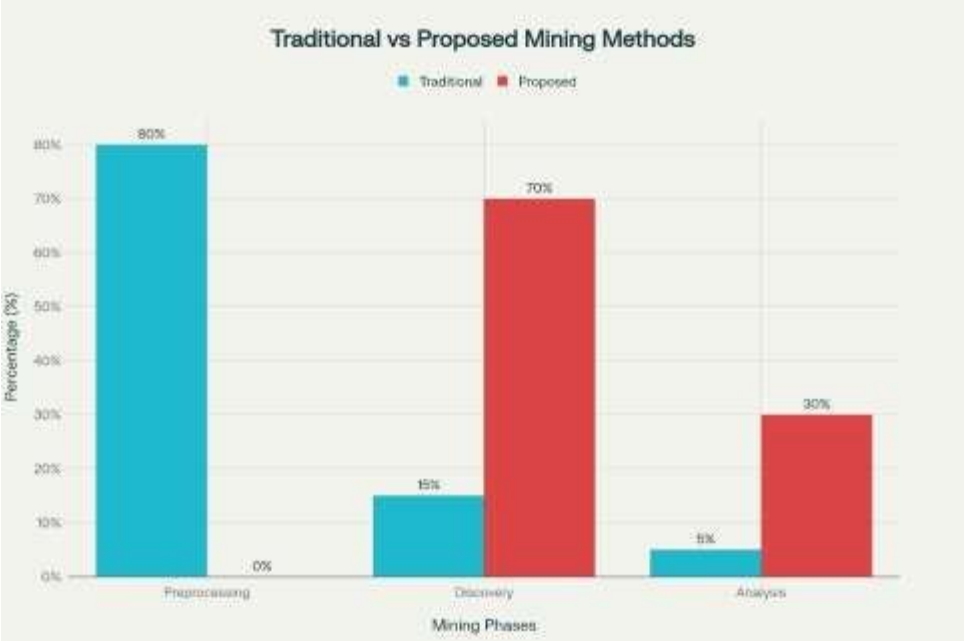

Fig 1. Comparison of time distribution across web usage mining phases between traditional server log analysis and the proposed server-side data collection method

Recent developments in privacy regulations, particularly the General Data Protection Regulation (GDPR), have further complicated traditional web analytics approaches.[2] Organizations face increasing pressure to maintain data sovereignty while ensuring compliance with stringent privacy requirements. Simultaneously, the demand for real-time analytics capabilities has grown substantially, requiring systems that can process and analyze data with minimal latency.

The methodology examined in this paper, originally presented by Canay and Kocabıçak [27], introduces a novel server-side approach that addresses these contemporary challenges through application-level data collection, eliminating preprocessing requirements, and providing comprehensive real-time analytics capabilities. This approach represents a paradigm shift from reactive log analysis to proactive data collection, fundamentally changing how organizations approach web usage mining.

Literature Review and Current Challenges

Traditional Web Usage Mining Preprocessing

The conventional web usage mining process consists of three primary phases: preprocessing, pattern discovery, and pattern analysis[5]. The preprocessing phase, which typically consumes the majority of processing time, involves several complex sub-tasks that have been extensively documented in the literature:

10.48047/jocaaa.2024.33.08.242

Data Cleaning: This initial step involves removing irrelevant entries from web server logs, including requests for static files (images, CSS, JavaScript), failed HTTP requests, and traces left by search engine crawlers. The manual nature of this process makes it both time-intensive and prone to inconsistencies.

User Identification: Traditional methods rely primarily on IP addresses and user agent strings to identify unique users, achieving only 70-80% accuracy due to issues such as dynamic IP assignment, proxy servers, and shared network access points. The mathematical formulation of this challenge demonstrates the complexity: given a set of IP addresses $IP = \{ip_1, ip_2, ..., ip_n\}$, browsers $B = \{b_1, b_2, ..., b_n\}$, and external links $K = \{k_1, k_2, ..., k_n\}$, determining unique users $U = \{u_1, u_2, ..., u_n\}$ from cleaned log entries becomes a computationally intensive probabilistic matching problem.[14]

Session Identification: This process involves segmenting user activities into discrete sessions, typically using time-based thresholds (commonly 30 minutes) or structure-oriented heuristics. However, these approaches face significant limitations when dealing with sessions that span midnight, dynamic IP changes during sessions, or users with extended idle periods.

Modern Privacy and Compliance Challenges

The implementation of GDPR and similar privacy regulations has fundamentally altered the web analytics landscape. Traditional client-side tracking methods face increasing restrictions due to ad-blockers, privacy-focused browsers, and user consent requirements. Studies indicate that up to 30% of users employ ad-blocking technologies, significantly impacting data collection accuracy.

Fig 2. Client-side vs server-side web tracking flow showing how Shopify stores send data through tag managers to third-party trackers and cookies, illustrating the ecosystem division between browser and server sides.[6]

10.48047/jocaaa.2024.33.08.242

Furthermore, data sovereignty concerns have intensified, particularly regarding the transfer of EU citizen data to US-based analytics platforms. The complexity of obtaining and managing user consent has created additional operational overhead for organizations seeking to maintain compliance while preserving analytical capabilities.

Real-Time Analytics Requirements

Modern digital experiences increasingly demand real-time analytics capabilities for immediate decision-making, personalization, and system optimization [8]. Traditional batch-processing approaches to web usage mining are inadequate for these requirements, as they introduce significant delays between data generation and actionable insights.

Real-time analytics systems must exhibit several key characteristics: low latency processing, high availability, scalability, and the ability to handle high-velocity data streams [21]. These requirements necessitate fundamentally different architectural approaches compared to traditional web usage mining methodologies.

Methodology: Server-Side Data Collection Architecture

Three-Tier System Design

The proposed methodology implements a comprehensive three-tier architecture that integrates data collection directly into the application framework, eliminating the need for separate preprocessing operations:

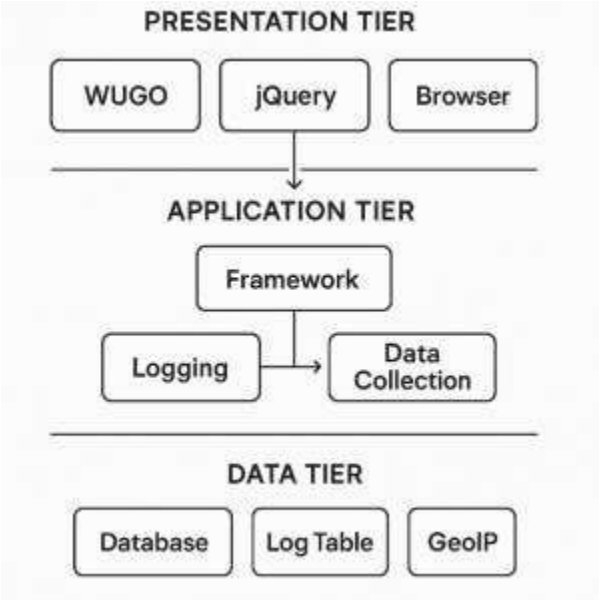

Fig 3. Three-tier architecture of the proposed server-side web usage mining data collection system

10.48047/jocaaa.2024.33.08.242

Presentation Tier: This layer encompasses traditional client-side web technologies including HTML, CSS, and JavaScript. Unlike conventional analytics implementations that rely heavily on client-side tracking scripts, this tier's primary function is user interface delivery with minimal tracking overhead.

Application Tier: The core innovation lies in this middle tier, which houses the integrated logging API. This server-side component operates as part of the web application framework, automatically capturing comprehensive usage data during normal request processing [17]. The API performs real-time data collection, cleaning, processing, and validation without requiring separate preprocessing steps.

Data Tier: This layer utilizes relational database systems optimized for Online Transaction Processing (OLTP), storing structured data in real-time. The design includes specialized tables for session management (log_session), page-level tracking (log_page), active session monitoring (open_sessions), user information (user_info), and geographic data (log_geoip).

Data Collection Process

The server-side data collection process operates through a systematic approach that captures comprehensive usage information during each HTTP request:

Automatic Data Capture: The logging API intercepts every page request at the server level, collecting data from multiple sources simultaneously[20]: HTTP request headers, network protocol information, application-specific user data, and external geographic location data derived from IP addresses.

Real-Time Processing: Unlike traditional methods that require offline batch processing, the proposed system performs data cleaning, validation, and storage operations immediately upon request receipt. This eliminates the temporal gap between data generation and availability for analysis.

Session Management: The system implements sophisticated session management through server-side session variables combined with database-backed tracking. This approach provides 100% accuracy in session identification, even when users access the system from dynamic IP addresses or through proxy servers.

Data Model and Storage Architecture

The proposed data model employs a normalized relational structure designed for both operational efficiency and analytical flexibility:

```
-- Core session data structure
log_session: {
   session_id, user_id, ip_address, user_agent,
   operating_system, browser_type, referrer_url,
   geographic_location, session_start_time
```

10.48047/jocaaa.2024.33.08.242

```
}

-- Detailed page-level tracking
log_page: {
   log_id, session_id, user_id, timestamp,
   requested_url, response_status, page_load_time,
   application_data, request_method, response_size
}
```

This structure enables complex analytical queries without requiring data transformation or restructuring, supporting both real-time monitoring and historical analysis.

Experimental Analysis and Performance Evaluation

Data Quality Assessment

Analysis of the implementation at Sakarya University demonstrated significant improvements in data quality across multiple dimensions. Over a 24-hour testing period, the system successfully processed 22,104 sessions and 161,672 pageviews with complete data integrity.

User Identification Accuracy: The proposed method achieved 100% accuracy in user identification through authenticated session management, compared to 70-80% accuracy typically achieved by IP-based heuristics in traditional systems. This improvement directly translates to more reliable user behavior analysis and personalization capabilities.

Session Integrity: Server-managed sessions eliminated common problems associated with traditional sessionization methods, including midnight session boundaries, proxy server interference, and cache-related path completion issues. The system successfully tracked sessions spanning multiple days without artificial termination.

10.48047/jocaaa.2024.33.08.242

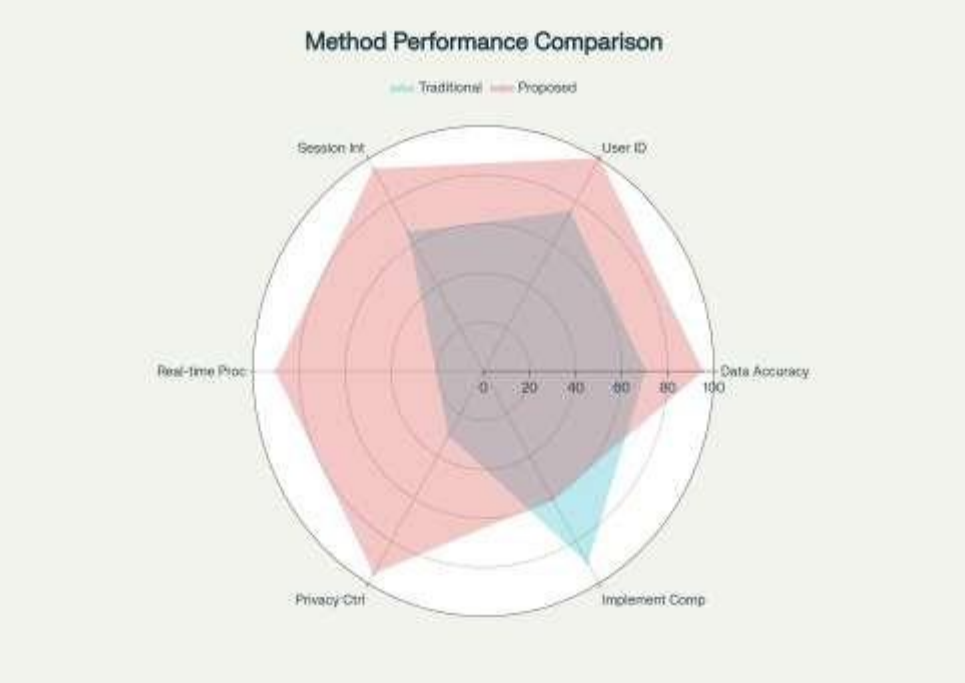

Fig 4. Performance comparison between traditional web server log analysis and the proposed server-side data collection method across key metrics

Performance Metrics and Efficiency Gains

The elimination of preprocessing overhead resulted in dramatic efficiency improvements in the overall web usage mining workflow[21]. Traditional approaches typically allocate 60-80% of processing time to preprocessing activities, while the proposed method redirects this effort toward pattern discovery and analysis.

Time Distribution Analysis: The proposed methodology enables organizations to allocate 70% of their analytical resources to pattern discovery and 30% to pattern analysis, compared to the traditional distribution of 80% preprocessing, 15% pattern discovery, and 5% pattern analysis.

Real-Time Processing Capabilities: The system demonstrated the ability to provide immediate insights into user behavior, supporting real-time decision-making processes that were previously impossible with batch-oriented traditional methods.

Privacy and Compliance Advantages

The server-side approach provides significant advantages for privacy compliance and data sovereignty:

10.48047/jocaaa.2024.33.08.242

Data Localization: All user data remains within the organization's infrastructure, eliminating concerns about international data transfers and third-party data processing. This approach directly addresses GDPR requirements for data controller accountability and territorial application.

Reduced Client-Side Tracking: By minimizing dependence on client-side tracking technologies, the system reduces vulnerability to ad-blockers and privacy-focused browser features that can compromise data collection accuracy.

Enhanced Security: Server-side processing enables comprehensive access logging and anomaly detection capabilities, supporting both security monitoring and regulatory compliance requirements.

Comparative Analysis with Contemporary Approaches

Traditional Server Log Analysis

Conventional web server log analysis suffers from several fundamental limitations that the proposed methodology addresses:

Data Incompleteness: Standard web server logs capture only basic request information, lacking application-specific context and user authentication details[26]. This limitation necessitates complex heuristic approaches for user and session identification, reducing overall accuracy.

Processing Overhead: The requirement for extensive preprocessing creates significant computational and temporal overhead, delaying the availability of analytical insights. Organizations must invest substantial resources in data cleaning and formatting before meaningful analysis can begin.

Limited Real-Time Capabilities: Traditional batch-processing approaches are incompatible with modern real-time analytics requirements, preventing immediate response to user behavior patterns.

Client-Side Analytics Solutions

Contemporary client-side analytics platforms, while popular due to their ease of implementation, face increasing limitations in the current privacy-conscious environment:

Privacy Regulation Challenges: GDPR and similar regulations have created complex consent management requirements for client-side tracking, often resulting in incomplete data collection.

Technical Limitations: Ad-blockers, tracking protection features, and privacy-focused browsers can prevent client-side tracking scripts from executing, creating significant gaps in data collection.

10.48047/jocaaa.2024.33.08.242

Performance Impact: Multiple client-side tracking scripts can negatively impact website performance, particularly on mobile devices with limited computational resources.

Hybrid and Server-Side Solutions

The proposed methodology aligns with emerging trends toward server-side analytics but offers several distinct advantages over existing hybrid approaches:

Complete Server-Side Processing: Unlike hybrid solutions that still rely partially on client-side data collection, the proposed method operates entirely on the server side, eliminating client-side dependencies.

Integrated Architecture: Rather than requiring separate analytics platforms or services, the proposed method integrates directly into the application framework, reducing complexity and improving performance.

Implementation Considerations and Scalability

Technical Requirements

Successful implementation of the proposed methodology requires careful consideration of several technical factors:

Database Performance: The system requires a robust OLTP database capable of handling high-frequency insert operations. Performance testing demonstrated the ability to process insert rates of 50-200 MB/s, supporting up to 1 million rows per second in typical use cases [13].

Storage Management: Long-term data retention requires strategic planning for data archiving and historical analysis. The system supports migration to OLAP structures or data warehouses for historical analytical processing.

Server Resources: The additional processing load from real-time data collection must be balanced against server capacity. The proposed architecture demonstrates linear scalability, with resource requirements directly proportional to traffic volume.

Scalability Architecture

The methodology supports horizontal scaling through several approaches:

Load Balancing: Multiple application servers can implement the logging API simultaneously, with centralized database storage ensuring data consistency.

10.48047/jocaaa.2024.33.08.242

Database Clustering: High-volume implementations can utilize database clustering or sharding techniques to distribute storage and processing loads.

Microservices Integration: The logging API can be implemented as a microservice, enabling independent scaling and deployment.

Future Research Directions and Limitations

Current Limitations

While the proposed methodology offers significant advantages, several limitations require acknowledgment:

Implementation Complexity: Unlike simple client-side analytics implementations, the proposed approach requires substantial programming expertise and database management capabilities.

Client-Side Data Limitations: Certain types of user interaction data, such as mouse movements, scroll behavior, and screen resolution, cannot be captured through server-side methods alone.

Initial Development Investment: Organizations must invest in custom development and system integration, contrasting with the immediate availability of commercial analytics platforms.

Future Enhancement Opportunities

Several areas present opportunities for methodology enhancement and research extension:

Machine Learning Integration: The structured, high-quality data produced by this methodology provides an excellent foundation for advanced machine learning applications, including predictive analytics, anomaly detection, and personalization algorithms [23].

Multi-Platform Extension: Future research could explore extending the methodology to mobile applications, IoT devices, and other digital interaction platforms beyond traditional web applications.

Advanced Analytics Applications: The elimination of preprocessing overhead creates opportunities for more sophisticated pattern discovery techniques, including real-time clustering, association rule mining, and sequential pattern analysis.

Conclusions and Implications

The server-side data collection methodology examined in this research represents a significant advancement in web usage mining capabilities, addressing fundamental limitations of traditional approaches while meeting contemporary requirements for privacy compliance and real-time analytics. The elimination of preprocessing

10.48047/jocaaa.2024.33.08.242

overhead, achievement of 100% user identification accuracy, and provision of comprehensive real-time monitoring capabilities demonstrate the practical value of this approach for modern digital organizations.

The methodology's emphasis on data sovereignty and privacy compliance positions it as a viable solution for organizations operating under stringent regulatory requirements, particularly in sectors such as healthcare, finance, and education where data sensitivity is paramount [25]. The complete elimination of client-side tracking dependencies addresses growing concerns about ad-blocker interference and privacy-focused browser restrictions.

From a practical implementation perspective, organizations considering this methodology must weigh the substantial initial development investment against the long-term benefits of superior data quality, real-time analytics capabilities, and regulatory compliance assurance [27]. The scalable architecture and direct database storage approach provide a foundation for advanced analytical applications that extend far beyond traditional web usage mining scenarios.

The research contributes to the broader evolution of web analytics toward more privacy-conscious, technically sophisticated, and organizationally controlled approaches. As digital privacy regulations continue to strengthen and user expectations for data protection grow, methodologies that prioritize data sovereignty while maintaining analytical capability will become increasingly valuable.

Future developments in this field should focus on reducing implementation complexity through standardized frameworks, extending the methodology to emerging platforms and interaction modalities, and exploring advanced applications of the high-quality data produced by server-side collection approaches. The foundation established by this methodology provides a robust platform for continued innovation in privacy-compliant, high- performance web analytics systems.

References

[1]Jain, V., & Kashyap, K. (2021). An efficient algorithm for web log data preprocessing. In *Machine Vision and Augmented Intelligence—Theory and Applications* (pp. 505–514). Springer, Singapore. https://doi.org/10.1007/978-981-16-5078-9_41

[2]Abdalla, A., Ahmed, T., & Seliaman, M. (2016). Web usage mining and the challenge of big data: A review of emerging tools and techniques. In I. R. M. Association (Ed.), *Big Data: Concepts, Methodologies, Tools, and Applications* (Vol. 6, Ch. 42, pp. 899–928). IGI Global. https://doi.org/10.4018/978-1-4666-9840-6.ch042

[3]Kumar, V., & Ogunmola, G. (2020). Web analytics for knowledge creation: A systematic review of tools, techniques, and practices. *International Journal of Cyber Behavior, Psychology and Learning, 10*(1), 1–14. https://doi.org/10.4018/IJCBPL.2020010101

10.48047/jocaaa.2024.33.08.242

[4]Čegan, L., & Filip, P. (2017). Webalyt: Open web analytics platform. In *2017 27th International Conference Radioelektronika* (pp. 1–5). IEEE. ttps://doi.org/10.1109/RADIOELEK.2017.7937605

[5]Tao, Y., Guo, S., Shi, C., & Chu, D. (2019). User behavior analysis by cross-domain log data fusion. *IEEE Access, 8*, 400–406. https://doi.org/10.1109/ACCESS.2019.2961769

[6]Ehikioya, S. A., & Lu, S. (2019). A path analysis model for effective e-commerce transactions. *African Journal of Computing & ICT, 12*(2), 55–71.

[7]Roy, R., & Rao, G. A. (2020). Survey on pre-processing web log files in web usage mining. *International Journal of Advanced Science & Technology*, *29*(3), 682–691.

[8]Ibrahim, K. K., & Obaid, A. J. (2021). Web mining techniques and technologies: A landscape view. *Journal of Physics: Conference Series, 1879*(3). https://doi.org/10.1088/1742-6596/1879/3/032125

[9]Srivastava, M., Srivastava, A. K., Garg, R., & Mishra, P. (2021). Performance evaluation of the MapReduce-based parallel data preprocessing algorithm in web usage mining with robot detection approaches. *IETE Technical Review*, 1–15. https://doi.org/10.1080/02564602.2021.1918584

[10]Bayir, M. A., & Toroslu, I. H. (2022). Maximal paths recipe for constructing web user sessions. *World Wide Web*, 1–31. https://doi.org/10.1007/s11280-022-01024-3

[11]Munk, M., & Benko, L. (2018). Using entropy in web usage data preprocessing. *Entropy*, *20*(1), 67. https://doi.org/10.3390/e20010067

[12]Srivastava, M., Garg, R., & Mishra, P. (2018). A MapReduce-based user identification algorithm in web usage mining. *International Journal of Information Technology and Web Engineering, 13*(2), 11–23. https://doi.org/10.4018/IJITWE.2018040102

[13]Knight-Davis, S. (2017). Using AWStats to analyze logs from EZProxy and from the public OPAC logs. In *Spring Forum: Collection Management and Technical Services Committees* (p. 228).

[14]Gamalielsson, J., Lundell, B., Butler, S., Brax, C., Persson, T., Mattsson, A., Gustavsson, T., Feist, J., & Lönroth, E. (2021). Towards open government through open-source software for web analytics: The case of Matomo. *JeDEM - eJournal of eDemocracy and Open Government*, *13*(2), 133–153. https://doi.org/10.29379/jedem.v13i2.650

[15]Aartsen, B., El-Gayar, O. F., & Noteboom, C. (2020). A systematic review of web usage mining techniques and future research options. *MWAIS 2020 Proceedings, 25*, 1–6.

[16]Milosevic, B., Regodic, D., & Saso, V. (2021). Big data management processes in business intelligence systems. In *Economic and Social Development: Book of Proceedings* (pp. 182–192). Varazdin Development and Entrepreneurship Agency.

10.48047/jocaaa.2024.33.08.242

[17] Zheng, G., & Peltsverger, S. (2015). Web analytics overview. In *Encyclopedia of Information Science and Technology* (3rd ed., pp. 7674–7683). IGI Global. https://doi.org/10.4018/978-1-4666-5888-2.ch756

[18] Srivastava, M., Garg, R., & Mishra, P. K. (2015). Analysis of data extraction and data cleaning in web usage mining. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)* (pp. 1–6). ACM. https://doi.org/10.1145/2743065.2743078

[19] Mishra, S., & Srivastava, S. (2021). Web development frameworks and its performance analysis – A review. *Smart Computing*, 337–343. https://doi.org/10.1201/9781003167488-39

[20] Onder, I., & Berbekova, A. (2021). Web analytics: More than website performance evaluation? *International Journal of Tourism Cities*. https://doi.org/10.1108/IJTC-03-2021-0039

[21] Gaurav Pandey. (2024). Decoding Digital Conversations: A Hybrid Sentiment Analysis Framework for WhatsApp Chat Behavioral Intelligence. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(21s), 4953 –. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/7503

[22] Shah, P., & Pandit, H. (2022). A review: Web content mining techniques. *Data Engineering & Smart Systems*, 159–172. https://doi.org/10.1007/978-981-16-2641-8_15

[23] Tyagi, N., & Gupta, S. K. (2018). Web structure mining algorithms: A survey. In *Big Data Analytics* (pp. 305–317). Springer. https://doi.org/10.1007/978-981-10-6620-7_30

[24] Lim, Z. Y., Ong, L. Y., & Leow, M. C. (2021). A review on clustering techniques: Creating better user experience for online roadshow. *Future Internet, 13*(9), 233. https://doi.org/10.3390/fi13090233

[25] Das, R., & Turkoglu, I. (2009). Creating meaningful data from weblogs for improving the impressiveness of a website by using path analysis method. *Expert Systems with Applications, 36*(3), 6635–6644. https://doi.org/10.1016/j.eswa.2008.08.067

[26] Manchanda, M., & Gupta, N. (2018). Web usage mining: Dynamic methodology to preprocessing web logs. *HELIX, 8*(5), 3810–3815. https://doi.org/10.29042/2018-3810-3815

[27] Canay, Ö., & Kocabıçak, Ü. (2023). An innovative data collection method to eliminate the preprocessing phase in web usage mining. *Engineering Science and Technology, an International Journal*. https://doi.org/10.1016/j.jestch.2023.101360