Augmented Web Usage Mining and User Experience Optimization with CAWAL's Enriched Analytics Data

Özkan Canay[1*] **and Ümit Kocabıçak**[2]

[1a]**Institute of Natural Sciences, Sakarya University, Sakarya, Türkiye**
[1b]**Vocational School of Sakarya, Sakarya University of Applied Sciences, Sakarya, Türkiye**
[2a]**Faculty of Computer and IT Engineering, Institute of Natural Sciences, Sakarya University, Sakarya, Türkiye**
[2b]**Turkish Higher Education Quality Council, Ankara, Türkiye**

Abstract

Understanding user behavior on the web is increasingly critical for optimizing user experience (UX). This study introduces Augmented Web Usage Mining (AWUM), a novel methodology designed to enhance web usage mining and improve UX. The approach is based on enriching the interaction data provided by CAWAL (Combined Application Log and Web Analytics), a web analytics model and framework, and utilizing these datasets in web mining. In this research, more than 1.2 million session records, collected over one month by CAWAL, were processed and transformed into 8.5 GB of enriched data. These datasets were analyzed using AWUM to explore session structures, page requests, service interactions, and exit methods. The results showed that 87.16% of sessions involved multiple pages, contributing to 98.05% of total pageviews. Moreover, 40% of users accessed various services, while 50% opted for secure exits, especially when dealing with personal or sensitive information. Association rule mining revealed key patterns in frequently accessed services, offering valuable insights into service integration and user preferences. These findings reveal CAWAL's superiority over conventional methods in precision and efficiency, providing a more comprehensive understanding of user behavior. AWUM demonstrates strong potential for advancing web mining in large-scale, multi-server environments and supports the development of more effective UX strategies through detailed interaction analysis.

Keywords

Web usage mining (WUM); augmented WUM; user experience (UX); behavior analysis; enriched analytics data

Paper type Research paper

* Corresponding author (canay@subu.edu.tr)

1 Introduction

Tracking visitor interactions and analyzing service usage are crucial for strategic decision-making processes in enterprise web applications and portals offering multiple services simultaneously. Web Usage Mining (WUM) analyzes user behavior and extracts meaningful patterns from this behavior to inform strategic decisions. These patterns help predict future user trends and ensure that decisions are based on solid foundations (Choudhary & Swami, 2023). Timely and accurate analyses provide valuable insights for developers and managers by offering concrete data to optimize web portal performance (Pastorino et al., 2019). However, understanding how users navigate between different services and interact with them in such complex systems becomes challenging with traditional methods, mainly because of large data sets and growing interaction diversity (Latha et al., 2023).

The conventional WUM process primarily relies on access logs from web servers, which typically contain basic information about which pages users access. However, these logs are insufficient for accurately defining user sessions and do not provide details on interactions within pages (Jin & Lin, 2022; Canay & Kocabicak, 2023). The limited structure of logs complicates the data analysis process, prolongs data processing times, and reduces the accuracy of analyses (Abílio et al., 2021). These limitations hinder efforts to improve the user experience and negatively affect strategic decision-making mechanisms (Husin et al., 2022). While various methods have been proposed in the literature to overcome these limitations, these methods have not been fully effective, mainly when applied to high-volume data.

Although the development of big data technologies and cloud-based solutions can potentially increase the efficiency of WUM processes, they present significant challenges regarding data security, privacy, cost, and performance. The incompatibility of traditional server logs with cloud systems complicates the data analysis, and real-time analysis performance can suffer (Mehrtak et al., 2021). Processing large-scale datasets in cloud environments increases costs, while slower data transfer speeds and processing delays limit performance. Furthermore, in decentralized cloud environments, data security and privacy become even more critical, and storing user data in such environments increases the risk of data breaches (Ageed et al., 2020; Mustafa et al., 2022). For these reasons, WUM processes involving extensive data sets face severe technical and financial barriers.

Several significant studies in the literature focus on WUM and user behavior tracking, mainly aimed at understanding user interactions within web applications. Research on web session clustering and user navigation behavior proposes methods for analyzing user movements and integrating this information into business development strategies (Bayir & Toroslu, 2022; Sharma & Malhotra, 2021). However, the inability of traditional WUM methods to accurately define sessions in large datasets and analyze some interactions reduces data quality, adversely affecting strategic decision-making processes (Husin et al., 2022). Additionally, studies on cloud-based solutions highlight issues related to data security, privacy, and costs, further driving the search for more effective solutions (Miller et al., 2022; Pang et al., 2023). Research on how WUM techniques can contribute to UX optimization emphasizes the importance of analyzing user interactions to improve UX and suggests developing strategies accordingly (Huidobro et al., 2022; Gayatri et al., 2022).

The CAWAL framework, proposed by Canay & Kocabicak (2024) and developed as a model combining application logs with web analytics, enables more comprehensive and accurate tracking of user interactions while providing an integrated solution to current methods. Unlike traditional approaches based on web server logs, this model allows the collection and integrated analysis of more extensive data at the application level, offering higher accuracy and performance in session identification and data processing. CAWAL also provides efficiency in data processing speed and session identification processes, standing out regarding data ownership and independence, making it a sustainable and cost-effective solution for large-scale organizations.

This study aims to use the extensive session and pageview data provided by the CAWAL framework, which has been processed through sessionization and data aggregation, to generate enriched analytical data. The Augmented Web Usage Mining (AWUM) approach, developed to utilize this data as a source for the WUM process, is designed to improve data quality and enhance the efficiency of analytical methods. This approach is expected to enhance resource use efficiency, speed, and result accuracy, providing more precise insights from large datasets than conventional WUM processes.

The research proposes that the enriched analytical data provided by the CAWAL framework and the AWUM approach will offer higher accuracy and performance than traditional WUM methods. The use of the CAWAL model in the WUM process and the effectiveness of the AWUM approach will be tested based on the following hypotheses:

H1: The enriched analytical data the CAWAL Framework provides offers higher data accuracy than traditional web server logs.

H2: The AWUM approach increases process efficiency by eliminating the pre-processing phase in traditional WUM processes.

H3: The enriched data provided by AWUM allows for a more in-depth analysis of user behavior, offering higher accuracy in optimizing the user experience (UX).

The capacity of the CAWAL framework to analyze user behavior is the primary focus of this study, and how these analyses can optimize web portals to meet user needs is examined. In this context, the framework's role in data collection and analysis, as well as its impact on user experience (UX), is evaluated in detail. Additionally, the advantages of CAWAL in terms of data security, performance, and cost-effectiveness are reviewed, with specific examples illustrating how these advantages can be applied to large-scale enterprise web portals.

The primary contributions of this study are as follows:

1. The AWUM approach was introduced, integrating enriched analytics data from the CAWAL framework for a more accurate and detailed analysis of user behavior compared to traditional methods.
2. The CAWAL framework simplifies the web usage mining process by removing the need for extensive pre-processing and delivering structured and high-quality data directly from application logs and web analytics.
3. CAWAL's enriched datasets improve the accuracy and efficiency of machine learning, user behavior prediction, and classification models.
4. The findings contribute to UX optimization by offering deep insights into user interactions, which can be used to enhance web portal performance and guide strategic decisions across various web services.

The structure of this paper is organized as follows. Section 2 reviews the literature on web usage mining and user experience, evaluating the limitations in addressing complex user behaviors and assessing previous work in these fields. Section 3 explains the application of the CAWAL framework in web usage mining through the AWUM approach. This section also addresses data reliability, privacy concerns, and the construction of enriched datasets. Section 4 presents four distinct, user experience-centered analyses focusing on user engagement, navigation patterns, exit methods, and service transitions to identify user behaviors and improve interaction with the web portal. Section 5 discusses the results, highlighting AWUM's advantages over traditional methods and their impact on user experience optimization. Finally, Section 6 concludes the paper by reflecting on the implications of the findings and proposing directions for future research.

2 Related work

Web Usage Mining is a process aimed at understanding user behaviors by analyzing web access logs and deriving meaningful patterns from this behavior. While web analytics typically focuses only on data collection and visualization, WUM extends the analysis by incorporating data cleaning, transformation, and extracting knowledge-driven patterns (Canay & Kocabicak, 2023). WUM processes typically begin with the processing of existing data rather than its collection, and their effectiveness largely depends on thorough data pre-processing. WUM studies in the literature generally rely on web server logs and focus on developing methodologies to enhance the quality of these logs (Yau & Zainon, 2020; Kumar & Gupta, 2022).

2.1 Web usage mining process and methods

The success of WUM processes is directly linked to the accuracy of the data pre-processing phase. Traditional server logs provide semi-structured data requiring extensive cleaning and transformation steps to produce meaningful insights. Cleaning raw data and removing noise are critical steps that enhance the reliability of the analysis. A. K. Srivastava & Srivastava (2023) have developed scalable data pre-processing methods, including heuristic techniques for robot detection, aiming to improve this stage significantly. Kaur & Garg (2019) emphasized the importance and effectiveness of techniques designed to enhance web data quality. However, these methods tend to be time-consuming and resource-intensive, particularly when applied to large datasets. Ali et al. (2020) proposed a comprehensive framework for pre-processing data to model user behaviors, underlining the importance of transforming raw web log data into analyzable patterns. Moreover, Asadianfam et al. (2020) demonstrated that integrating case-based reasoning and clustering techniques into WUM processes enables a deeper analysis of user behaviors.

The discovery phase of pattern recognition is crucial in identifying user behavior patterns. Despite the challenges in pre-processing, traditional WUM methods have succeeded in clustering users based on behavior patterns, as demonstrated by Singh & Kaur (2021), who found clustering algorithms effective in grouping users with similar behavior patterns. Ouf et al. (2023) successfully improved customer recommendation systems' accuracy by combining different mining techniques. These analyses focus on the ability to predict future user behaviors. Munk et al. (2021) emphasized that WUM data allow proactive design changes by forecasting user behaviors. Roy & Rao (2022) proposed an efficient framework for web usage mining based on time and fairness constraints to analyze user behaviors and provide personalized recommendations promptly and fairly. Malik et al. (2021) enhanced classification accuracy by combining random forest algorithm results with ant colony optimization (ACO). Serin et al. (2022) investigated a fuzzy C-means-based reduced feature set association rule mining approach to predict web user behavior patterns. Similarly, Elsheweikh (2023) proposed a new web recommendation model based on web usage mining techniques.

WUM methods have evolved significantly in recent years, particularly with the integration of machine learning algorithms, which have enhanced the accuracy of user behavior analysis. For instance, Asadianfam et al. (2020) employed case-based reasoning techniques to group users and analyze behaviors, while Singh & Kaur (2021) utilized clustering techniques for the same purpose. Both approaches have proven effective, yet challenges persist in managing the scale and complexity of modern web applications. In addition, WUM-based frameworks are widely applied in e-commerce websites to increase customer retention and satisfaction. For example, Waqas et al. (2018) demonstrated the value added to businesses through the analysis of web log files, which allowed the identification of user behavior patterns. Furthermore, Cahaya & Siswanti (2020) explored the impact of Internet banking service quality on e-customer satisfaction and loyalty, emphasizing the importance of WUM-based applications in enhancing customer satisfaction.

2.2 The role of web usage mining in user experience design

User experience (UX) design aims to optimize websites based on user expectations and focuses on enhancing users' interactions with the site. Incorporating UX design principles into web platforms is essential for improving usability and engagement. User-centered design (UCD) plays a crucial role in this process and is vital for successful UX outcomes (Lallemand et al., 2015). UX optimization follows continuous improvement through A/B testing, usability testing, and feedback surveys. A/B testing compares the performance of different design alternatives, while usability testing analyzes users' interactions with the system. The literature emphasizes that these tests provide concrete feedback to improve design decisions by enhancing the effectiveness of UX (Sowbhagya et al., 2023).

Web usage mining (WUM) and user experience (UX) design complement each other, working together to improve user experiences by analyzing user behaviors. WUM provides UX designers valuable data for proactive design changes by analyzing users' online behaviors (Ali et al., 2021). For instance, WUM data on e-commerce platforms enables personalized product recommendations by analyzing users' purchasing patterns. This personalization enhances customer satisfaction while strengthening users' engagement with the platform (Benali, 2022). These studies demonstrate that incorporating WUM insights into UX strategies is essential for improving user engagement and satisfaction, particularly in large-scale web environments.

In addition to enhancing user satisfaction, WUM also facilitates personalized recommendations and plays an essential role in identifying user pain points and abandonment patterns. The importance of using interaction data to develop predictive models for future user behavior is frequently emphasized in the literature (Wasino et al., 2023). Such models empower UX designers to identify challenges and implement appropriate design improvements quickly. A successful example in this regard is the Cluster-N-Engage framework proposed by Lim, Ong, & Leow (2023), which utilizes WUM data to uncover significant opportunities for UX enhancement.

Recent studies further demonstrate the critical role of WUM in enhancing UX by offering insights into user interactions and behaviors across various platforms. Table 1 provides an overview of these studies, highlighting their key contributions, methods, and techniques from the past four years, summarizing the latest advancements in WUM and UX integration.

Table 1: Recent studies related to WUM and UX fields.

Key Contribution	Methods & Techniques Used	Reference
Categorizes users based on navigation patterns in online directories to enhance user segmentation.	Association Rule Mining, Apriori, Fuzzy clustering	Athinarayanan et al. (2023)
Reveals user requirements by discovering patterns from web log data through web usage mining.	Data Mining, Pattern Analysis, Clustering, Classification	Dubey et al. (2024)
Develops a collaborative recommendation system model for improving user experience.	Clustering, Rule Extraction, Neural Network, Genetic Algorithm	Elsheweikh (2023)
Enhances web interfaces by increasing the predictability of user interactions using web usage mining.	Clustering, Session Reconstruction Algorithm, Bayesian Network	Jörs & De Luca (2023)
Discovers user navigation patterns through session clustering and measures user engagement.	Clustering, K-Means, K-Medoids, Bisecting K-Means	Lim, Ong, Leow, Lee, & Tay (2023)
Develops personalized recommendation systems based on user shopping preferences in e-commerce.	Recommendation Systems, Apriori Algorithm	Mahesh Kumar & Om Prakash (2021)
Improves user experience on a government website by analyzing user behavior.	Association Rule Mining, Apriori, FP-Growth, Sequential Pattern Mining	Rawira & Esichaikul (2023)
Understands user behavior and provides personalized web experiences through recommendation systems.	Clustering, Association Rule Mining, K-Means, Fuzzy C-Means	Serin et al. (2022)
Analyzes user behavior in e-commerce websites using web usage mining for personalization.	Association Rule Mining, Apriori Algorithm	Soewito & Johan (2022)
Proposes an innovative method for user profile creation and updating to improve user experiences.	User Profiling, Cognitive Psychometric Memory Model	Sowbhagya et al. (2023)
Presents a new recommendation system to improve website structure by shortening user navigation paths.	Recommendation Sys., Reinforcement Learning, Adaptive ranking	Ting et al. (2024)
Identifies user interests on e-commerce sites and provides personalized recommendations.	Clustering, K-Means, DBSCAN, Hybrid Density-Based K-Means	Win & Lwin (2024)

2.3 Challenges in existing WUM tools and UX integration

Many existing WUM tools still rely on traditional web server logs, which often fail to capture the full range of user interactions, creating significant data gaps in UX optimization (Menezes & Nonnecke, 2014). These logs typically lack details on specific activities and in-page behaviors, limiting the scope of analysis. Ouf et al. (2023) emphasizes that addressing these shortcomings requires more detailed data collection and advanced analytical methods. In response, M. Srivastava et al. (2022) introduced and evaluated preprocessing techniques, including robot detection, within their MapReduce-based parallel data preprocessing algorithm to improve the efficiency of web usage mining.

Despite recent progress, current research does not thoroughly address several critical WUM and UX optimization aspects. The integration of WUM with UX design often remains superficial, particularly when examining specific details of user interactions within applications (Sharma & Malhotra, 2021). Additionally, few studies have explored the analysis of dynamic in-page behaviors or how such data can be incorporated into UX optimization processes, which remains a significant gap (Rawira & Esichaikul, 2023).

Furthermore, there is a notable lack of research on data diversity and real-time analysis. Most studies continue to focus on static datasets, paying insufficient attention to the real-time analysis of streaming data and its implications for UX optimization (Kumar et al., 2022). Machine learning and artificial intelligence techniques in WUM and UX integration are also underdeveloped, leaving considerable untapped potential for proactive UX improvements based on behavior prediction (Xing-hai, 2023). Moreover, concerns about data privacy and security are becoming increasingly prominent, requiring the development of new methods to process and analyze user data securely (Zagan & Danubianu, 2023).

The CAWAL framework, central to this study, addresses many of these limitations by integrating web analytics with application-level logs and providing enriched datasets that capture complex user interactions across multiple services. This approach enhances the accuracy of user behavior tracking and facilitates real-time processing of high-dimensional data, which traditional WUM tools struggle to handle efficiently. CAWAL's sessionization and cross-service tracking capabilities offer the necessary infrastructure for detailed user behavior analysis, particularly in multi-service portals and multi-server web farms, where interactions are more complex and fragmented.

3 Methodology

The CAWAL (Combined Application Log and Web Analytics) model and its software framework (Canay & Kocabicak, 2024) implementation were integrated into CAWIS (Campus Automation Web Information System), Sakarya University's large-scale institutional web application (Canay et al., 2011). The CAWIS system, based on a portal architecture, operates within a load-balanced web farm and uses separate subdomains for each service. The interaction data collected from the CAWIS and its use in WUM processes are explained in depth, including how CAWAL eliminates the data pre-processing phase, in the following sub-sections.

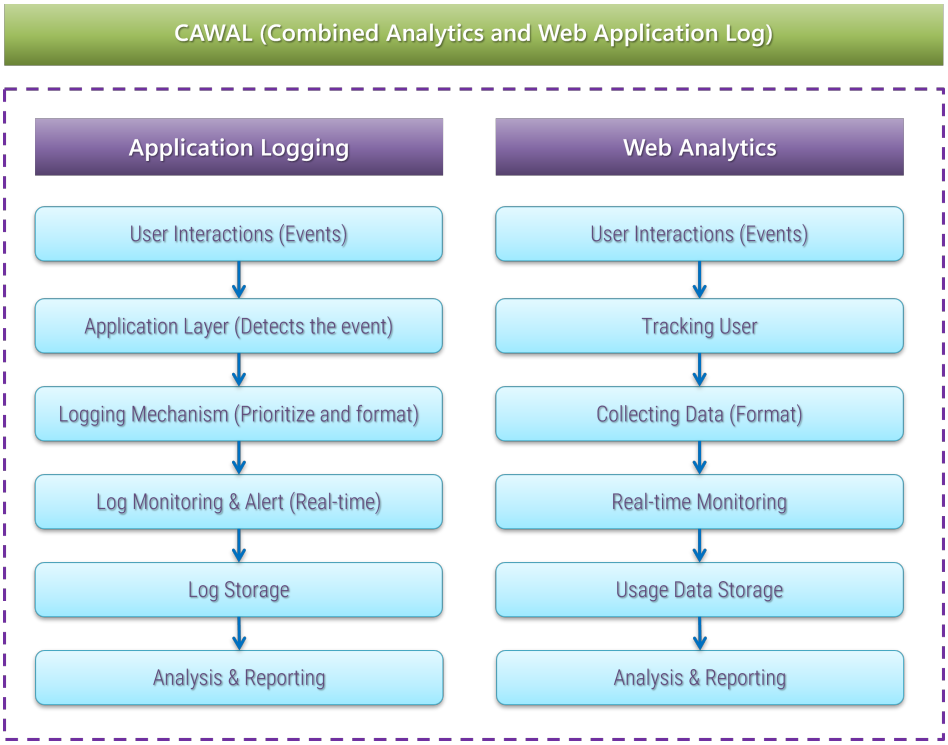

Figure 1: Application logging and web analytics processes combined by CAWAL.

3.1 CAWAL model and framework

The CAWAL model combines application logging with web analytics and extends the specialized data collection methodology, proposed previously (Canay & Kocabicak, 2023), to a broader perspective, addressing enterprise web portals and multi-server architectures. The analytical software framework developed as a practical application of this model includes an API for data collection, a database model for data storage, and a method for generating analytical insights. This framework enables the comprehensive collection of page movement data on web portals, facilitating its use in data mining and business intelligence applications.

CAWAL was developed based on traditional software development practices and later expanded to include web analytics features. Figure 1 illustrates the fundamental characteristics shared between application logging and web analytics. While conventional web analytics emerged to capture basic user interactions, modern tools now focus on gathering comprehensive data relevant to applications. In contrast, the CAWAL model was designed primarily to log application activities, with web analytics as a secondary function. This dual capability distinguishes CAWAL from other analytics tools by not only collecting detailed application logs but also integrating them into advanced web usage analytics processes.

3.2 Integrating CAWAL into web usage mining

WUM utilizes rich data sources to understand users' interactions with websites and to enhance user experience based on these interactions. The CAWAL framework systematically consolidates and processes these data sources, playing a crucial role in utilizing web usage data for analytical insights. The quality of the data used in mining processes is a critical factor that directly influences the accuracy and efficiency of the analysis (Wang & Li, 2023). Server logs, the traditional source of information for WUM, often contain raw, semi-structured data. This type of data presents significant challenges in the preprocessing phase, which is one of the first and most critical steps in WUM (Choudhary & Swami, 2023).

The pre-processing stage, which includes data cleaning, transformation, and user and session identification, varies based on the project, dataset, and methods employed. It typically constitutes a substantial portion of the time and effort spent on a WUM activity (Raman & Raj, 2021). The pre-processing stage, carried out entirely offline, involves complex operations and requires significant effort. However, it often fails to produce precise and successful results due to data quality, technical limitations, data diversity and complexity, pre-processing methods, and human involvement (Prakash et al., 2021).

However, the session and page navigation data collected by the CAWAL model, supported with application data, are inherently accurate and well-structured. This comprehensive data set provides the necessary information for WUM and serves as a unique and high-quality data source. Figure 2 illustrates how the data generated by the CAWAL model is optimized for use as a data source in WUM. This approach effectively eliminates the most critical and challenging step of the WUM process—preprocessing—while maximizing the process's success and enabling more detailed and accurate tracking of user interactions.

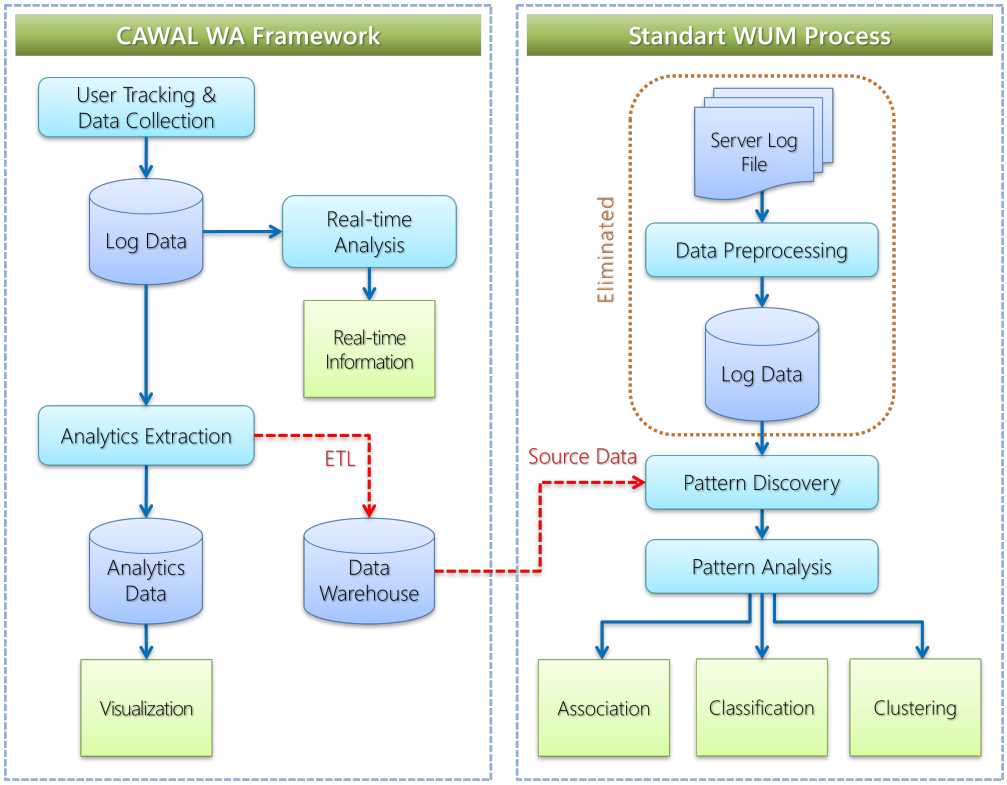

Figure 2: Usage of web access data collected with CAWAL as data source in web usage mining.

3.3 Augmented web usage mining approach

Using high-quality, accurate raw analytical data stored in CAWAL's data warehouse eliminates the need for traditional server logs in web usage mining and renders the pre-processing step unnecessary (Canay & Kocabicak, 2023). This innovative approach accelerates the analysis process by reducing the time and resources required for data cleaning, normalization, and transformation of the data collected by CAWAL, thereby improving the accuracy of the data mining results. Additionally, adopting a structured data format facilitates the understanding and the analysis of complex and challenging user behavior patterns. Figure 3 compares the standard WUM process, as described by M. Srivastava et al. (2019), with the WUM process implemented through the CAWAL framework.

The success of web usage mining is directly related to the quality of the raw data. In a standard WUM workflow, the procedure begins with cleaning server log data, identifying users, and determining session identities, followed by various techniques such as path analysis, trend analysis, and classification. However, the traditional method is constrained by the quality of the raw data, which may result in a loss of accuracy in the final analyses. The CAWAL framework handle these limitations through a series of data augmentation steps that enhance the richness and accuracy of the data before entering the WUM pipeline.

This process, termed Augmented Web Usage Mining (AWUM) in this study, begins with session augmentation, where connected tables, browser data, and user data are integrated to provide a more comprehensive view of user sessions and activities. It is followed by user activity augmentation, which enriches session and pageview information with detailed data such as login and logout information, server identification, and page duration. These steps culminate in enriched web analytics data, referred to as accurate data, which provides the foundation for more reliable and in-depth analyses when incorporated into the WUM process. This innovative approach enables more robust results in advanced clustering, association rule mining, and predictive modeling, allowing for a better understanding of user behavior and web usage patterns.

3.4 Data reliability and privacy

The absence of synthetic data in this study guarantees the reliability and integrity of the findings. All data were collected from the now-retired CAWIS web portal, developed by Sakarya University, through the CAWAL framework. The data collection process adhered to the Internet Services Usage Policy Agreement approved by the users and was fully compliant with the university's regulations. Additionally, all necessary permissions for the study were obtained from Sakarya University, ensuring compliance with both the university's internal policies and the country's legal framework.

Diverse data anonymization techniques were applied, including the irreversible masking of user identities and the shifting of timestamps, to further safeguard user privacy. These methods effectively prevent the re-identification of user data and enhance data security. Furthermore, these adjustments aim to enhance user anonymity and reduce the risk of potential data breaches. The timestamp modification is not expected to negatively impact the analysis results, as the study focuses on trends and behavioral patterns within a specific timeframe. Ev-

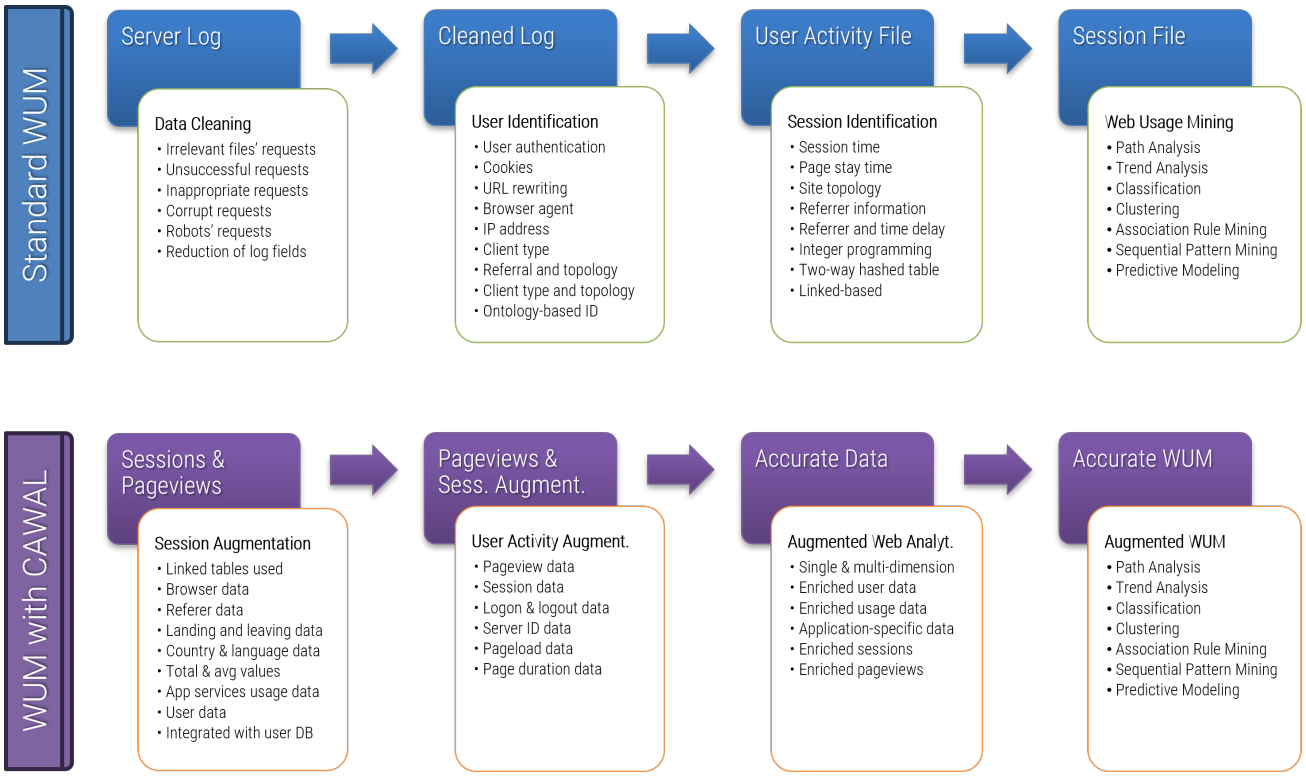

Figure 3: Standard WUM process and CAWAL's Augmented WUM.

ery research step has been carefully conducted to safeguard participant privacy and ensure data security.

3.5 Generation of enriched analytics data

Web usage mining processes often require the handling of large and complex datasets. A comprehensive process was carried out to transform the detailed raw session and user data collected according to the data model by the CAWAL framework into enriched analytics data following sessionization procedures. Data selection and enrichment were crucial steps in preparing the data stored in the data warehouse for WUM applications. Sophisticated SQL queries created for the enriched session and pageview table views were foundational in processing data sets effectively and extracting meaningful insights.

Enriched session and pageview data were obtained through complex SQL queries performed on data marts spanning one month and totaling 8.5GB. These extensive queries, which allow for the efficient querying, filtering, and transformation of large and heterogeneous data sets, are critical for producing advanced analytics results and multi-dimensional data sets. These enriched data sets have been used to analyze user behaviors, gain a deeper understanding of user needs, and enhance the overall efficiency of the web portal's usage.

3.6 Schemas of data files

After the sessionization process, the enriched session data includes not only the data retrieved from the session table for each record but also summary statistics, such as the first, last, total, and average values, obtained through SQL queries. These statistics are accompanied by user-related information and pageview data corresponding to the session. The field names, descriptions, and sample data of the CSV files, which contain only numerical and categorical session information for use in web usage mining processes, are presented in Table 2. This schema includes critical metrics such as session duration and page loads, forming the core of the data used in WUM processes.

In addition to the session data, fields summarizing user interactions with the portal services during each session are also included. Fields starting with "s_" (e.g., s_gate, s_mail, s_obis) indicate whether the user visited the respective service, with "1" representing a visit and "0" indicating no visit. Fields starting with "p_" (e.g., p_gate, p_mail, p_obis) record the number of pages the user navigated within each service. Lastly, fields beginning with "r_" (e.g., r_gate, r_mail, r_obis) represent the ratio of the number of pages visited within each service to the total number of pages visited during the session.

These fields provide detailed numerical and categorical data capturing various aspects of the sessions, which

Table 2: Schema of the CSV file containing enriched numerical session data.

Field Name	Description	Sample Data
Log_ID	Log ID number.	12359285
Session_ID	Session ID number.	83665107
Log_Date_Time	Log date and time information.	22.11.2022 13:00
User_ID	User ID number.	184922
Session_Login_Status	Login status in the session.	1
Logins_During_Period	Number of logins during the period.	16
User_Type	User type (e.g. Acd./Adm./Stud./Unit).	6
Sex	User gender (Undefined/Male/Female).	2
Age	User age	18
Age_Group	User age group (Categorical, 1-4).	1
User_Language_TR	Browser language.	1
User_Location	IP location.	1
Browser_Type	Web browser type.	1
Referer_Type	Reference type.	6
Landing_Srv_ID	Service ID number first accessed.	7
Exit_Srv_ID	Service ID number logged out.	3
Exit_Type	Logout type.	0
Total_Session_Duration	Total session duration.	730
Avg_Page_Duration	Average page duration.	48.67
Total_Page_Load	Total page load (generation) time.	2.88
Avg_Page_Load	Average page load (generation) time.	0.19
Page_Count	Total number of pages navigated in a session.	15
Visitor_PageView	Total number of pages navigated as a visitor.	2
User_PageView	Total number of pages navigated as a user.	13
Service_Count	Total number of services browsed.	2
Page_per_Service	Average number of pages per service.	7.5
Visited_Service_IDs	ID numbers of the visited services.	“1,3”

Table 3: CSV data files and their properties where enriched session and pageview data are stored.

Time Frame	Time Range	File Name .CSV	Record Count	File Size (MB)
1-week	2022-11-21 - 2022-11-27	va_page4	3,158,694	266.00
1-month	2022-11-01 - 2022-11-30	va_sess5	1,220,916	235.00

are structured for use in WUM processes. The metrics detailing session characteristics are crucial for accurately analyzing user behavior, allowing a comprehensive understanding of session structures. The enriched session data is ideal for analyzing users' overall navigation behavior throughout a session, summarizing key elements such as session duration, page loads, and visited services. This dataset plays a critical role in macro-level analyses, such as identifying patterns in session duration and service visits.

Similarly, pageview data contains detailed information about the pages visited during each user session alongside session-specific data. As a critical data source for web usage mining processes, these records include details such as the date and time of each page visit, the duration of time spent on the page, the page load time, and the page's unique identifier. Pageview data is also enriched with demographic information retrieved from the session table, including user type, gender, age group, browser type, and IP location. The reciprocal relationship between session and pageview data defines the terms used to describe these datasets, while the resulting enriched data schemas enhance the effectiveness of WUM activities conducted on it.

3.7 Preparation of time-frame-based datasets

As a result of the executed queries, all data related to sessions and pageviews were meticulously collected to analyze complex relationships and reveal multi-layered data structures. This data was organized through views created in the database. The data, reprocessed to include numerical and categorical information for WUM activities, was saved as separate views. In the final stage, the data corresponding to the periods specified in Table 3 was extracted from these views and transformed into enriched session and pageview datasets. These datasets, amounting to 8.5 GB collected over a one-month period, were structured and stored in CSV format to facilitate easy processing using data analysis tools such as Python.

After being collected through the CAWAL framework, the enriched session and pageview data play a pivotal role in analyzing user behavior over time. The enriched session data capture all actions performed by users during their sessions, including essential metrics such as session duration, the number of pages visited, and login-logout activities. For instance, the session data collected over one month comprises 1,220,916 records, while a one-week pageview dataset includes 3,158,694 records. The enriched pageview data provide detailed information on each visit, including metrics like pageview duration, load time, and user login status. These datasets are optimized to enable in-depth analysis of user behavior over specific time frames. Storing the data in CSV format facilitates quick and efficient processing, ensuring high accessibility for web usage mining applications.

4 User experience-driven analysis and results

Four analyses were conducted on enriched analytics data via the Augmented Web Usage Mining (AWUM) approach to enhance the efficiency of the web portal and optimize the user experience. These analyses were performed using Python programming language to address user engagement, experience, portal navigation, and interaction.

The first analysis examines the frequency of user exits and their correlation with specific user attributes, aiming to identify exit patterns across different user profiles. The second analysis focuses on the methods users employ to leave the system, identifying disruptions in user experience and proposing improvements to mitigate these issues. The third analysis evaluates the transition paths between portal services and their impact on user experience. The fourth and final analysis uncovers behavioral patterns by extracting association rules from user interactions.

These analyses provide detailed modeling of user behaviors, such as session transitions, pageviews, and service usage patterns, employing web usage mining techniques, including pattern analysis, user segmentation, behavioral modeling, service transition analysis, and association rule mining.

4.1 Analysis of bounce rate and client attributes

Bounce rate, a significant web analytics metric, represents the percentage of sessions where users view only a single page on the website and leave without any further interaction. This rate is a critical indicator for understanding users' engagement levels with the content and the impact of website design on user experience. The detailed statistical analysis of the monthly dataset reveals the differences between single-page and multi-page sessions, focusing on metrics such as the number of sessions and pageviews.

Out of the total 1,220,916 sessions in the examined dataset, 156,707 are single-page sessions, while 1,064,209 are multi-page sessions. Single-page sessions constitute 12.84% of total sessions, while multi-page sessions make up 87.16%. For pageviews, single-page ses-

Table 4: Distribution of client attributes according to sessions with single and multiple-page visits.

Client Attributes (Categories)	Single Page	Multiple Pages
Browser_Type:		
1-Standard Browser	47.55%	99.17%
2-Search Engine	14.63%	0.01%
3-Text-Based Browser	37.82%	0.82%
Referer_Type:		
1-From Homepage	1.88%	7.62%
2-From Portal Services	20.32%	15.44%
3-From Corporate Domains	4.02%	16.95%
4-From Search Engine	4.23%	24.44%
5-Other Referrers	0.31%	1.25%
6-No Referrer	69.24%	34.29%
User_Language_TR:		
0-Other	1.53%	2.47%
1-Turkish	98.47%	97.53%
User_Location:		
0-In Türkiye	21.37%	19.61%
1-Internal (SAU)	77.56%	79.51%
2-Outside Türkiye	1.07%	0.87%

sions directly correspond to the session count (156,707 pageviews), while multi-page sessions generate a total of 7,882,632 pageviews. This analysis indicates that 1.95% of the total pageviews are from single-page sessions, and 98.05% are from multi-page sessions.

These findings provide significant insights into user engagement levels and their connection with the content on the site, offering critical data for evaluating the website's user experience. The distribution of the four categorical attributes used to distinguish users during their server access and their distribution between single-page and multi-page user groups is presented in detail in Table 4.

The data shows that the number of single-page sessions and the page impressions generated by these sessions are much lower than the number of multi-page sessions and page impressions. These findings indicate that multi-page sessions involve significantly more interaction and content consumption on the website. According to the analyzed data, standard browsers have a high utilization rate for multi-page viewing sessions. This observation suggests that users of standard browsers engage more with the content and interact more actively with the site. Additionally, the bounce rate is higher for visits from text-based browsers and search engines, primarily due to search engine indexing robots and automated site review tools, commonly known as web crawlers. Since these browsers do not support cookies, their actions are recorded as single-page sessions within the system. However, this group should also consider portal services that meet user needs on a single page, such as the menu service, and the immediate closure of any portal page in case of accidental opening.

The significant proportion of single-page viewing sessions without a redirector indicates that users arrive at the site by directly entering URLs or using bookmarks. Many sessions originating from search engines and involving multiple pageviews show that users accessed the site via search engines rather than using bookmarks or directly typing the site name. After finding the content they were looking for, users continued to browse the site, demonstrating the effectiveness of the SEO efforts. As expected, the web portal is predominantly used by Turkish-speaking users based on the client's browser language. This observation confirms that the site's content optimization in the local language is successful and meets the needs of local users. Additionally, the location distribution reveals a predominance of in-house and in-country users, suggesting that the site is more relevant for geographically close users.

The chi-squared (χ^2) test results from the detailed analysis of single-page and multi-page session data indicate statistically significant differences in the distributions across various categories. In this test, the χ^2 statistic is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i represents the observed frequency, and E_i is the expected frequency for each category. The degree of freedom is determined by the formula:

$$\text{Degrees of Freedom (DoF)} = (r-1)(c-1)$$

where r is the number of rows and c is the number of columns in the contingency table. A p-value of less than 0.05 indicates a statistically significant relationship between client attributes and the bounce rate. Table 5 presents the results of this analysis, showing the impact of client characteristics on the bounce rate across four categories.

The test results show a statistically significant difference between single-page and multi-page sessions' distribution of browser type and reference type. The high Chi-square values and low p-values (<0.05) for these two attributes indicate a significant relationship with the type of session (single-page vs. multi-page), suggesting that the observed differences are unlikely to be due to random variation. On the other hand, the low Chi-square values and high p-values for the user language (Turkish) and user location attributes suggest no significant relationship with the type of session, and any slight differences that exist may be attributable to random variation. The degrees of freedom (DoF) are calculated as the number of categorical levels of the tested attribute minus one since

Table 5: Effects of client attributes on bounce rate.

Client Attributes	Chi2	p-value	DoF
Browser_Type	68.19	0.00	2
Referer_Type	38.72	0.00	5
User_Language_TR	0.00	1.00	1
User_Location	0.12	0.94	2

the analysis focuses on a single variable. For instance, for Browser_Type, which has three categories (Standard Browser, Search Engine, Text-Based Browser), the DoF is derived as 3 - 1 = 2. The DoF, alongside the Chi-square value and p-value, is crucial in determining the statistical significance of the relationship.

Bounce rate is often linked to the quality of a website's initial impression and the relevance of its content, commonly seen as an indicator of user engagement. While a high bounce rate is frequently viewed negatively, it does not always suggest a poor user experience. For example, if users quickly find what they need, such as through a service menu, their immediate exit may reflect successful navigation rather than disengagement. However, when high bounce rates are concentrated on specific pages or audience segments, it can indicate areas where the user experience or content requires improvement. Understanding these patterns is essential for identifying the key factors driving user behavior and provides a solid foundation for optimizing site design and content strategies.

4.2 Service-based analysis of exit methods

The ideal way for users to leave a web application after completing their tasks is to log out using the "secure exit" method before exiting the system. However, due to the nature of the web, it is impossible to compel users to log out securely if they choose not to. Failing to log out securely may lead to security vulnerabilities and compromise the accuracy of the collected analytical data.

The exit method analysis is mathematically modeled using a directed graph $G = (V, E)$, where each service is represented as a node $v \in V$, and each transition between services, including exit methods, is represented as a directed edge$e = (v_i, v_j) \in E$. The weight $w_{v_i \to v_j}$ of each edge corresponds to the frequency of transitions from service v_i to service v_j, capturing user behavior and exit preferences across the web portal. This graph allows for analyzing both service transitions and preferred exit methods.

In the graph, each node's degree $d(v)$ is determined by the sum of incoming and outgoing transitions, representing the activity level of each service concerning user exit strategies. The edge weights $w_{v_i \to v_j}$, normalized to ensure comparability, provide insights into the most common paths users take when leaving services, highlighting the relative importance of secure exits versus direct exits such as closing the browser without logging out.

Figure 4, based on the analysis of one week's session data, presents a network visualization of the methods users prefer when exiting various services on the web portal. This visualization, representing an essential data set, provides a basis for discussing user behavior trends and exit strategies on the portal and for strategic decisions to enhance the user experience.

The two primary methods for exiting the portal are "regular session termination" via the secure exit button and "direct system exit" without logging out. A small subset of session transactions, handled through the warning window, is included in the regular session termination group. Both primary methods, represented by the pur-

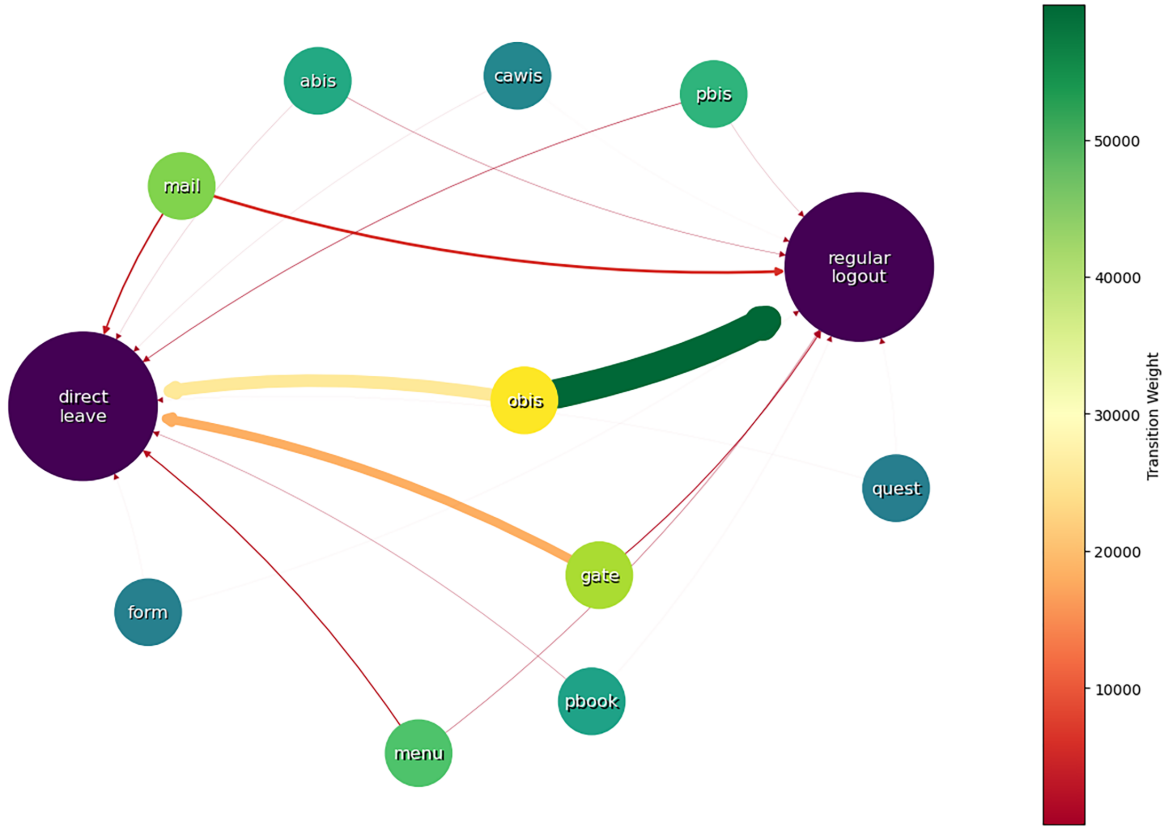

Figure 4: Service-based analysis of portal exit methods.

ple color indicating preference intensity, are used with approximately equal frequency. The color and thickness of the edges (connections) between services and exit methods represent transition weights. These transition weights are visualized by a color scale ranging from green to red, with red indicating higher frequency and greater intensity of transitions.

When portal services examine exit preferences, using services such as "gate," "obis," "mail," and "pbis" as exit points are more pronounced, with thicker, more vivid edges indicating greater intensity. This observation indicates that users generally prefer to log out when exiting these services securely. Possible reasons for this include the critical nature of services that handle personal information, privacy concerns, and increased security awareness.

The analysis reveals that users mostly prefer to directly close the window or navigate to another address when exiting the "gate" service, while they are more cautious to securely log out when leaving services containing personal information, such as "obis" and "mail." The preference for secure exit in the "obis" service, which caters to students, is expected, given that some students use computers in shared institutional spaces.

Analyzing user behavior offers crucial insights for optimizing web portal design and improving user interaction. For instance, the predominant use of the secure exit option indicates that users are consciously terminating their sessions, suggesting that this function should remain easily accessible. Additionally, the significant use of the direct exit option indicates that session timeout durations or automatic logout functions should be reviewed to enhance security further.

4.3 User interaction analysis through portal service transitions

Analyzing user transitions between services provides valuable information about how users navigate through different sections of the web portal. Examining these interaction patterns makes it possible to identify which services are frequently accessed together and how users move between them during their sessions. Understanding these transitions is crucial for improving the overall user experience, as it can highlight potential bottlenecks, points of friction, or popular pathways users tend to follow.

Additionally, this analysis helps uncover how effectively the portal's services are integrated and whether users encounter difficulties when switching between services. By evaluating these patterns, organizations can make informed decisions about where to focus their efforts in improving service accessibility and flow, ultimately enhancing the portal's usability. Figure 5 presents a network visualization of user transitions between services in the web portal. The transitions and their corresponding weights highlight the most common navigation paths, providing a clear overview of user behavior within the portal.

The data reveal which services are most frequently utilized, as indicated by the higher transition weights be-

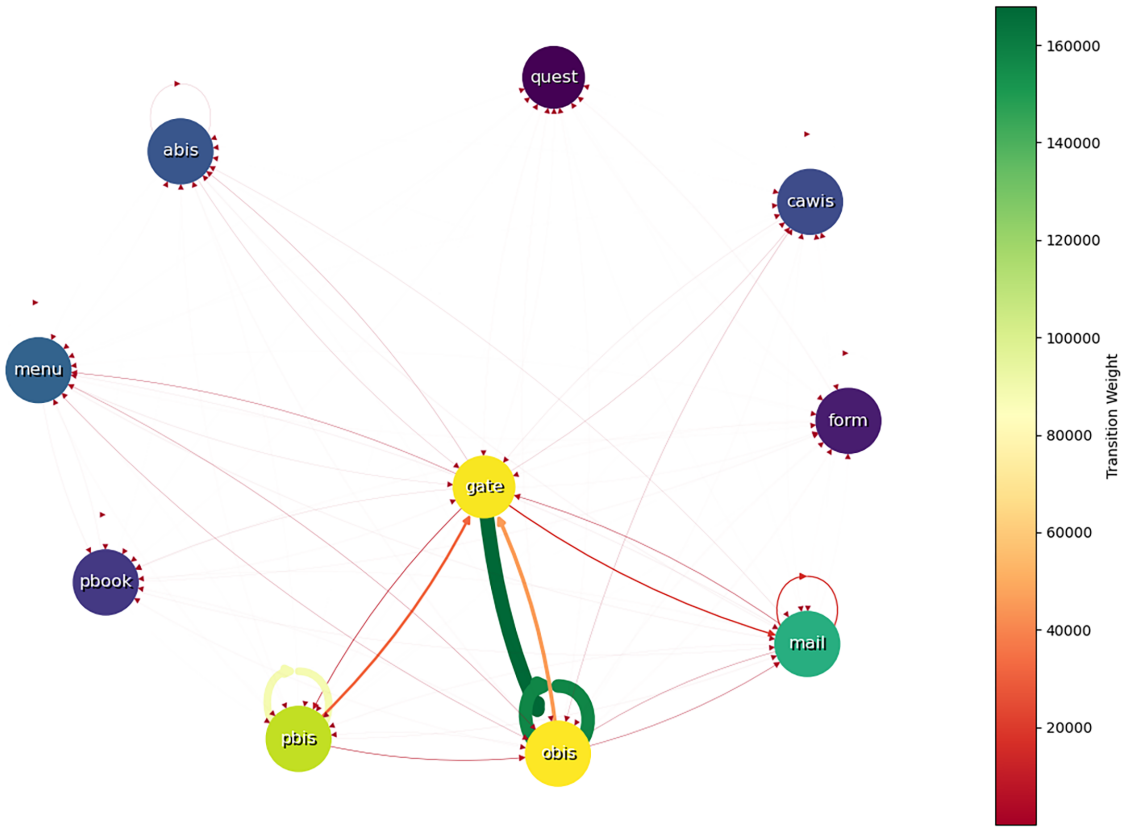

Figure 5: Transition paths and weights between portal services.

tween certain services. For instance, the central node "gate" is connected to many edges, indicating that it is the most frequently used service within the portal, functioning as the primary entry point into the system. In contrast, services such as "pbis," "obis," and "mail," which are also heavily used, are represented by bright nodes, while clear lines depict the transition weights to other services. The representation of bright nodes and clear lines indicates substantial user activity between these services and others.

The visualization also shows services like "cawis," "quest," and "form," characterized by darker nodes and fewer transitions. These patterns suggest that these services are utilized by a smaller subset of users, likely for specific tasks. The self-connecting arc above certain services, particularly the "gate" service, represents transitions from those services back to the system login, emphasizing the cyclical nature of user behavior in accessing and re-accessing the system.

4.4 Association rule mining on service interaction data

Association rule mining conducted on service interaction data aims to identify patterns in user behavior and interactions to uncover services frequently used together and understand how these relationships influence the overall user experience. This analysis provides a solid foundation for designing both software interfaces and service content by revealing co-used services and their impact on various user segments.

The Apriori algorithm, one of the most commonly used methods in association rule mining, was employed in this analysis. Apriori works by iteratively identifying frequent itemsets, where the support of an itemset is the proportion of transactions that contain the itemset. The support for an itemset A is calculated as:

$$\text{Support}(A) = \frac{\text{Number of transactions containing } A}{\text{Total number of transactions}}$$

After identifying frequent itemsets, association rules are generated based on these sets. An association rule is in the form of $A \rightarrow B$, where A and B are itemsets. The confidence of an association rule is the conditional probability

Table 6: Top 30 association rules derived using Apriori.

Antecedents	Consequences	Antecedent Support	Conclusion Support	Support	Confidence	Lift	Leverage	Conviction	Zhangs Metric
abis	gate	0.456	0.961	0.454	0.994	1.034	0.015	6.843	0.061
cawis	gate	0.405	0.961	0.405	1.000	1.040	0.016	infinite	0.065
form	gate	0.376	0.961	0.376	1.000	1.040	0.014	infinite	0.062
form	mail	0.376	0.843	0.374	0.993	1.178	0.057	22.954	0.243
menu	gate	0.474	0.961	0.474	1.000	1.040	0.018	infinite	0.074
pbook	gate	0.433	0.961	0.433	1.000	1.040	0.017	infinite	0.068
quest	gate	0.322	0.961	0.322	1.000	1.040	0.013	infinite	0.057
menu	mail	0.474	0.843	0.467	0.984	1.167	0.067	9.643	0.273
pbook	mail	0.433	0.843	0.430	0.994	1.180	0.066	26.412	0.268
quest	mail	0.322	0.843	0.322	1.000	1.187	0.051	infinite	0.232
cawis, abis	gate	0.302	0.961	0.302	1.000	1.040	0.012	infinite	0.055
cawis, abis	mail	0.302	0.843	0.299	0.992	1.176	0.045	18.394	0.215
form, abis	gate	0.276	0.961	0.276	1.000	1.040	0.011	infinite	0.053
form, abis	mail	0.276	0.843	0.276	1.000	1.187	0.043	infinite	0.217
abis, gate	mail	0.454	0.843	0.446	0.983	1.166	0.064	9.223	0.261
mail, abis	gate	0.446	0.961	0.446	1.000	1.040	0.017	infinite	0.070
abis, menu	gate	0.330	0.961	0.330	1.000	1.040	0.013	infinite	0.058
obis, abis	gate	0.322	0.961	0.322	1.000	1.040	0.013	infinite	0.057
pbis, abis	gate	0.400	0.961	0.397	0.994	1.034	0.013	5.992	0.054
pbook, abis	gate	0.327	0.961	0.327	1.000	1.040	0.013	infinite	0.058
abis, menu	mail	0.330	0.843	0.330	1.000	1.187	0.052	infinite	0.235
obis, abis	mail	0.322	0.843	0.320	0.992	1.177	0.048	19.652	0.222
pbis, abis	mail	0.400	0.843	0.394	0.987	1.171	0.058	12.184	0.244
pbook, abis	mail	0.327	0.843	0.327	1.000	1.187	0.052	infinite	0.234
form, cawis	gate	0.281	0.961	0.281	1.000	1.040	0.011	infinite	0.054
form, cawis	mail	0.281	0.843	0.281	1.000	1.187	0.044	infinite	0.219
form, cawis	menu	0.281	0.474	0.276	0.982	2.070	0.143	28.655	0.719
mail, cawis	gate	0.387	0.961	0.387	1.000	1.040	0.015	infinite	0.063
cawis, menu	gate	0.340	0.961	0.340	1.000	1.040	0.013	infinite	0.059
obis, cawis	gate	0.343	0.961	0.343	1.000	1.040	0.013	infinite	0.059

that a transaction containing A also contains B, defined as:

$$\text{Confidence}\,(A \rightarrow B) = \frac{\text{Support}\,(A \cup B)}{\text{Support}\,(A)}$$

Lift is another crucial metric used to evaluate the strength of an association rule, indicating how much more likely B is to occur given A compared to its general occurrence. It is defined as:

$$\text{Lift}\,(A \rightarrow B) = \frac{\text{Confidence}\,(A \rightarrow B)}{\text{Support}\,(B)}$$

A lift value greater than 1 indicates a strong positive association between A and B, while a value less than 1 suggests that A and B are negatively correlated.

The analysis was conducted using the "va_page4.csv" dataset, which contains detailed records of user interactions with various services on the web portal. The "Service_ID" column was first mapped to meaningful English names for easier identification of each service, as proper labeling enhances the interpretability of the results. Attributes such as user type, age group, location, browser type, and reference type were selected based on their potential influence on user interactions. These features facilitate a more precise examination of user navigation patterns and how diverse user segments interact with the portal. The service access order within each session was also determined to capture sequential patterns in user behavior, a critical step in understanding how users move through the portal's services.

The dataset was converted into transaction lists using the Transaction Encoder to facilitate the analysis. One-hot encoding was applied to prepare the data for machine learning algorithms by converting categorical data into binary format. This step ensures that the data is suitable for processing by the Apriori algorithm, which requires a binary matrix representation. The Apriori algorithm was used to identify frequent itemsets with a minimum support value of 25% and a confidence threshold of 98%. Table 6 presents the top thirty association rules, with lift values greater than or equal to 1, ensuring only the most significant and reliable rules are included.

The association rules outlined here reveal strong relationships between preceding services, such as "abis," "cawis," and "form," and subsequent services, such as "gate" and "mail." Metrics such as antecedent support, consequent support, support, confidence, leverage, lift, conviction, and Zhang's metric were analyzed in detail to understand users' tendencies to switch between services. For instance, the analysis found that most users who visited the "abis" service then navigated to the "gate" service, with a confidence level of 99.44%. The leverage value of 1.0343 indicates that this transition occurs 3.43% more frequently than would be expected by random chance. Similarly, the relationship between "form" and "mail" exhibited a notable connection, with a leverage value of 1.1784, indicating a higher-than-expected association.

The high conviction values of these rules suggest that the likelihood of the consequent occurring without the antecedent is extremely low. This finding is particularly highlighted by the infinite conviction values for transitions from "form" and "cawis" to "gate," indicating solid associations. Zhang's metric values were also evaluated to determine the reliability and direction of these relationships, confirming the robustness of these findings.

5 Discussion

Existing approaches in web usage mining (WUM) and user experience (UX) optimization face significant data accuracy and process efficiency limitations. The CAWAL framework employed in this study and the proposed AWUM method aim to overcome these limitations by providing an innovative solution for a deeper examination of user activity. The hypotheses of the study are based on the premise that the datasets provided by the CAWAL would offer more accurate and thorough assessments compared to conventional techniques and that these evaluations would provide strategic insights for optimizing UX.

Four critical analyses were conducted to test the hypotheses, specifically targeting different aspects of the web portal's engagements and user experience. The first analysis focused on testing the first hypothesis. The accuracy and analytical depth of the datasets provided by CAWAL were examined in detail based on users' session durations and page transitions on the web portal. The findings indicate that the framework provides more thorough and precise assessments than traditional methods. For example, out of 1,220,916 sessions, 156,707 of these, representing 12.84%, were single-page visits, often referred to as bounce visits, while multi-page sessions made up 98.05% of the total pageviews, demonstrating the framework's ability to analyze user activity more accurately. These results strongly support the validity of the first hypothesis.

The second and third analyses were conducted to test the second hypothesis, which assessed the CAWAL framework's potential to optimize UX. Similar to the efforts of Malik et al. (2021) to enhance random forest algorithms with ant colony optimization (ACO), this study applied a method that improves the accuracy of predictive models by enhancing data quality. In the service-based exit method examination, it was clearly observed that 50% of users preferred secure exit methods, while the remaining 50% predominantly used direct exit methods, such as closing the browser tab or window, navigating to another URL, or leaving the portal without logging out. The secure exit rate for services containing personal information, such as "obis" and "mail", rising to 75% demonstrates that secure exit options are already prioritized for these services. The investigation of transitions between portal services revealed that 85% of users used the "gate" service as the first entry point, with remarkable transitions occurring between other services. These findings clearly support the second hypothesis by illustrating the framework's ability to capture detailed user behavior across multiple services.

One of the major contributions of this study to the literature is the introduction of the Augmented Web Usage Mining (AWUM) approach. AWUM, developed using

the enriched analytical datasets provided by the CAWAL framework, offers a more extensive examination of user activity than conventional web log analysis. AWUM allows for a deeper understanding of the underlying dynamics of these engagements, moving beyond superficial analyses of user interactions with specific pages and services. These findings directly support the third hypothesis by demonstrating how enriched datasets facilitate a more detailed analysis of user behavior. For instance, user navigation patterns and interactions between frequently accessed services were examined, revealing significant insights into how specific service transitions, such as between "mail" and "obis," resulted in higher task completion rates. These insights show that AWUM's enriched data contributes to optimizing the user experience, validating the third hypothesis.

It is essential to note that the analyses in this study contribute to the validation of all three hypotheses in a comprehensive manner. For example, the enhanced data accuracy achieved through the CAWAL framework not only improves the precision of user behavior analysis (H3) but also streamlines the entire WUM process, making it more efficient (H2). The removal of the preprocessing phase through AWUM accelerates the data analysis workflow, reducing computational complexity, which contributes to both process efficiency (H2) and improved data accuracy (H1). Additionally, the enriched datasets provided by AWUM offer valuable insights for optimizing the user experience (H3) by revealing detailed patterns in user behavior and service transitions, allowing for more informed decisions in UX design. These combined improvements highlight the overall effectiveness of the methodologies used in this study and their potential to enhance multiple aspects of web usage mining.

This effectiveness is further demonstrated by the CAWAL framework's substantial advancements in data accuracy and processing efficiency when compared to other approaches. For instance, while the web recommendation model proposed by Elsheweikh (2023) focuses on surface-level analyses of user activity, CAWAL's enriched datasets enable a more detailed and in-depth understanding across multiple services. This advanced analytical capability provides deeper insights into user engagements, making a notable contribution to the existing body of literature. Additionally, integrating the fuzzy C-means-based association rule mining method proposed by Serin et al. (2022) with CAWAL could further optimize WUM processes, facilitating more informed decision-making for UX improvement and increased user satisfaction.

Furthermore, frameworks like the time and fairness-constrained WUM approach developed by Roy & Rao (2022) have made progress in developing personalized recommendation systems. However, CAWAL goes further by offering more extensive datasets and analytical capabilities, as demonstrated by the high accuracy and F1 scores achieved in predictive models. This advancement reinforces the reliability of CAWAL's assessments over traditional methods. Nonetheless, the dataset used in this study covers a specific period and user group, which may limit the generalizability of the findings.

Given these considerations, it is clear that the comprehensive data provided by the CAWAL framework enhances the precision and depth of user behavior analysis. The Augmented Web Usage Mining (AWUM) approach, developed based on this framework and supported by enriched analytical data, enables detailed examination not only of superficial user interactions but also of the underlying dynamics. This method demonstrates superior accuracy and analytical depth compared to existing approaches, positioning CAWAL as a valuable model for both academic research in the fields of WUM and UX, as well as practical applications in industries such as e-commerce, education, healthcare, and public services, where effective analysis of complex user behaviors is crucial.

6 Conclusion and future work

This study demonstrates that the CAWAL model significantly improves data accuracy and process efficiency in web usage mining, providing a clearer understanding of user behavior and enhancing user experience. By eliminating the time-consuming preprocessing phase, CAWAL speeds up the WUM process and reduces manual intervention. Analyses of over 1.2 million session records show that the framework offers a more precise and detailed representation of user activity, outperforming traditional methods in data accuracy. These improvements have directly enhanced the quality of user behavior analysis and service transitions.

A vital advantage of the CAWAL model is its ability to deliver practical findings through the Augmented Web Usage Mining (AWUM) approach. The enriched datasets allow for a thorough analysis of critical user behavior patterns, such as exit methods and service transitions. For instance, results show that 85% of users started their sessions at the portal gateway, while services handling personal data, such as "obis" and "mail," had a secure exit rate of 75%. These findings demonstrate the framework's capability to guide improvements in service integration and security features. Additionally, CAWAL ensures strong privacy protection through advanced data anonymization techniques, making it suitable for environments prioritizing data confidentiality.

Despite the promising outcomes, certain limitations must be addressed. This study relied on data from a specific period and user group, which may limit the generalizability of the findings. Future research should apply the CAWAL model to more extensive and more diverse datasets across different industries to assess its performance in various operational settings. Furthermore, exploring ways to improve data security, such as advanced encryption and secure data management, will be crucial for broader adoption in sensitive applications. Addressing these aspects will help further develop CAWAL and establish it as a valuable model for web usage mining and user experience optimization.

Ethics Statement

All data used in this study were collected from the CAWIS web portal in compliance with Sakarya University's regulations and the legal framework of the Republic of Turkey. Necessary permissions were obtained from Sakarya University, and diverse data anonymization techniques were applied to ensure user privacy and data security throughout the research process.

Consent Statement

Consent for data usage was obtained through the Internet Services Usage Policy Agreement, which was approved by all users of the CAWIS web portal.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Özkan Canay https://orcid.org/0000-0001-7539-6001
Umit Kocabıçak https://orcid.org/0000-0003-0369-9737

Data availability statement

The data are not publicly available due to privacy reasons and confidentiality agreement restrictions.

References

Abílio, R., Garcia, C. M., & Fernandes, V. (2021). Data mining applied on web robots detection: a systematic mapping. In *Anais do 15. congresso brasileiro de inteligência computacional*. http://doi.org/10.21528/cbic2021-60

Ageed, Z. S., Ibrahim, R. K., & Sadeeq, M. A. M. (2020). Unified ontology implementation of cloud computing for distributed systems. *Current Journal of Applied Science and Technology*, 82–97. http://doi.org/10.9734/cjast/2020/v39i3431039

Ali, N. M., Gadallah, A. M., Hefny, H. A., & Novikov, B. (2020). An integrated framework for web data preprocessing towards modeling user behavior. In *2020 international multi-conference on industrial engineering and modern technologies, fareastcon 2020*. http://doi.org/10.1109/FarEastCon50210.2020.9271467

Ali, N. M., Gadallah, A. M., Hefny, H. A., & Novikov, B. A. (2021). Online web navigation assistant. *Vestnik MGSU*, *15*(3), 351–364. http://doi.org/10.35634/VM210109

Asadianfam, S., Kolivand, H., & Asadianfam, S. (2020). A new approach for web usage mining using case-based reasoning and clustering techniques. *SN Applied Sciences*, *2*(7), 1251. http://doi.org/10.1007/s42452-020-3046-z

Athinarayanan, S., Joel, M. R., Jumlesha, S., Susmitha, K., & Charitha, L. (2023). Using pattern analysis and machine learning to categorise users of online directories based on their surfing habits. In *International conference on sustainable communication networks & applications, icscna 2023 - proceedings* (pp. 1089–1094). http://doi.org/10.1109/ICSCNA58489.2023.10370171

Bayir, M. A., & Toroslu, I. H. (2022). Maximal paths recipe for constructing web user sessions. *World Wide Web*, *25*(6), 2455–2485. http://doi.org/10.1007/s11280-022-01024-3

Benali, K. (2022). Ontology and web usage mining for website maintenance and user experience improvement. *International Journal of Data Mining, Modelling and Management*, *14*(4), 372–400. http://doi.org/10.1504/IJDMMDM.2022.126666

Cahaya, Y. F., & Siswanti, I. (2020). Internet banking service quality in indonesia and its impact on e-customer satisfaction and e-customer loyalty. *Management Research Studies Journal*, *1*(1), 75–83. http://doi.org/10.56174/mrsj.v1i1.350

Canay, O., & Kocabicak, U. (2023). An innovative data collection method to eliminate the preprocessing phase in web usage mining. *Engineering Science and Technology, an International Journal*, *40*. http://doi.org/10.1016/j.jestch.2023.101360

Canay, O., & Kocabicak, U. (2024). Cawal: A novel unified analytics framework for enterprise web applications and multi-server environments. *Information Processing and Management*, *61*(3). http://doi.org/10.1016/j.ipm.2023.103617

Canay, O., Meric, S., Evirgen, H., & Varan, M. (2011). Realization of campus automation web information system in context of service unity architecture. In *International symposium on computing in science & engineering (iscse)* (pp. 173–179). Izmir, Turkey.

Choudhary, L., & Swami, S. (2023). Exploring the landscape of web data mining: an in-depth research analysis. *Current Journal of Applied Science and Technology*, *42*(24), 32–42. http://doi.org/10.9734/cjast/2023/v42i244179

Dubey, S. M., Tiwari, G., & Narwaria, P. (2024). Server access pattern analysis based on weblogs classification methods. In *Emergent converging technologies and biomedical systems* (pp. 183–195). Springer Nature Singapore. http://doi.org/10.1007/978-981-99-8646-0_16

Elsheweikh, D. L. (2023). A novel web recommendation model based on the web usage mining technique. *Journal of Applied Information Technology*, *14*(5), 1019–1028. http://doi.org/10.12720/jait.14.5.1019-1028

Gayatri, M., Satheesh, P., & Rao, R. S. (2022). Deep learning for user behaviour prediction using streaming analytics. *International Journal on Recent and Innovation Trends in Computing and Communication*, *10*(2s), 289–297. http://doi.org/10.17762/ijritcc.v10i2s.5946

Huidobro, A., Monroy, R., & Cervantes, B. (2022). A high-level representation of the navigation behavior of website visitors. *Applied Sciences*, *12*(13), 6711. http://doi.org/10.3390/app12136711

Husin, H. S., Thom, J. A., & Zhang, X. (2022). Evolution of user navigation behavior for online news. *International Journal of Web Information Systems*, *18*(1), 1–22. http://doi.org/10.1108/ijwis-06-2021-0064

Jin, J., & Lin, X. (2022). Web log analysis and security assessment method based on data mining. *Computational Intelligence and Neuroscience*, *2022*, 1–9. http://doi.org/10.1155/2022/8485014

Jörs, J. M., & De Luca, E. W. (2023). Predictive behavior modeling through web graphs: Enhancing next page prediction using dynamic link repository. In *2023 ieee international conference on web intelligence and intelligent agent technology (wi-iat)* (pp. 415–420). http://doi.org/10.1109/WI-IAT59888.2023.00068

Kaur, J., & Garg, K. (2019). Efficient management of web data by applying web mining pre-processing methodologies. In *Advances in intelligent systems and computing* (Vol. 731, pp. 115–122). http://doi.org/10.1007/978-981-10-8848-3_11

Kumar, A., Bhushan, B., Pokhriya, N., Chaganti, R., & Nand, P. (2022). Web mining and web usage mining for various human-driven applications. In *Advanced practical approaches to web mining techniques and application* (pp. 138–162). IGI Global. http://doi.org/10.4018/978-1-7998-9426-1.ch007

Kumar, A., & Gupta, R. (2022). Enhancing data quality in web usage mining: A comprehensive review. *Journal of Information Systems Engineering & Management*. http://doi.org/10.5120/ijca2022919772

Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience: a concept without consensus? exploring practitioners' perspectives through an international survey. *Computers in Human Behavior*, *43*, 35–48. http://doi.org/10.1016/j.chb.2014.10.048

Latha, K., Sulaiman, E., & Yohannan, S. (2023). Linkage of digitalization and perceived organizational performance of small and medium enterprises. *SDMIMD Journal of Management*, 47–60. http://doi.org/10.18311/sdmimd/2023/32687

Lim, Z., Ong, L., & Leow, M. (2023). Cluster-n-engage: a new framework for measuring user engagement of website with user navigational behavior. *IEEE Access*, *11*, 112276–112292. http://doi.org/10.1109/access.2023.3322958

Lim, Z., Ong, L., Leow, M., Lee, T., & Tay, Q. (2023). Understanding user behaviour with web session clustering and user engagement metrics. In *2023 19th ieee international colloquium on signal processing & its applications (cspa)* (pp. 19–24). http://doi.org/10.1109/CSPA57446.2023.10087488

Mahesh Kumar, S., & Om Prakash, R. (2021). Knowledge-based recommendation system for online business using web usage mining. In *Rising threats in expert applications and solutions: Proceedings of ficr-teas 2020* (pp. 293–300). http://doi.org/10.1007/978-981-15-6014-9_34

Malik, V., Mittal, R., Singh, J., Rattan, V., & Mittal, A. (2021). Feature selection optimization using aco to improve the classification performance of web log data. In *2021 8th international conference on signal processing and integrated networks (spin)* (pp. 671–675). http://doi.org/10.1109/SPIN52536.2021.9566126

Mehrtak, M., SeyedAlinaghi, S., MohsseniPour, M., Noori, T., Shamsabadi, A., Heydari, M., & Dadras, O. (2021). Security challenges and solutions using healthcare cloud computing. *Journal of Medicine and Life*, *14*(4), 448–461. http://doi.org/10.25122/jml-2021-0100

Menezes, T. C., & Nonnecke, B. (2014). Ux-log: Understanding website usability through recreating users' experiences in logfiles. In *International journal of virtual worlds and human computer interaction*. http://doi.org/10.11159/vwhci.2014.006

Miller, K., Rosenberg, J., Pickard, O., Hawrusik, R., Karlage, A., & Weintraub, R. (2022). Segmenting clinicians' usage patterns of a digital health tool in resource-limited settings: Clickstream data analysis and survey study. *Journal of Medical Internet Research*, *24*(4), e30320. http://doi.org/10.2196/30320

Munk, M., Pilkova, A., Benko, L., & Blazekova, P. (2021). Methodology of stakeholders' behaviour modelling for data-driven ux design optimization. *MethodsX*, *8*, 101570. http://doi.org/10.1016/j.mex.2021.101570

Mustafa, R. A., Chyad, H. S., & Mutar, J. R. (2022). Enhancement in privacy preservation in cloud computing using apriori algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(3), 1747–1757. http://doi.org/10.11591/ijeecs.v26.i3.pp1747-1757

Ouf, S., Helmy, Y., & Ashraf, M. (2023). Web mining techniques - a framework to enhance e-commerce user experience. *International Journal of e-Collaboration*, *19*(1), 315790. http://doi.org/10.4018/IJeC.315790

Pang, P. C., Munsie, M., & Chang, S. (2023). A method for analyzing navigation flows of health website users seeking complex health information with google analytics. *Informatics*, *10*(4), 80. http://doi.org/10.3390/informatics10040080

Pastorino, R., Vito, C. D., Migliara, G., Glocker, K., Binenbaum, I., Ricciardi, W., & Boccia, S. (2019). Benefits and challenges of big data in healthcare: An overview of the european initiatives. *European Journal of Public Health*, *29*(3), 23–27. http://doi.org/10.1093/eurpub/ckz168

Prakash, P. G. O., Jaya, A., Ananthakumaran, S., & Ganesh, G. (2021). Predicting the user navigation pattern from web logs using weighted support approach. *Indonesian Journal of Electrical Engineering and Computer Science*, *21*(3), 1722–1730. http://doi.org/10.11591/ijeecs.v21.i3.pp1722-1730

Raman, G., & Raj, G. (2021). Mutual information pre-processing based broken-stick linear regression technique for web user behaviour pattern mining. *International Journal of Intelligent Engineering and Systems*, *14*(1), 244–256. http://doi.org/10.22266/ijies2021.0228.24

Rawira, P., & Esichaikul, V. (2023). Web usage mining for determining a website's usage pattern: A case study of government website. In *Communications in computer and information science* (Vol. 1942, pp. 88–100). http://doi.org/10.1007/978-981-99-7969-1_7

Roy, R., & Rao, G. A. (2022). A framework for an efficient recommendation system using time and fairness constraint based web usage mining technique. *International Journal of Innovation & Learning*, *27*(3), 298–315. http://doi.org/10.18280/isi.270308

Serin, J., SatheeshKumar, J., & Amudha, T. (2022). Efficient fuzzy c-means based reduced feature set association rule mining approach for predicting the user behavioral pattern in web usage mining. *Journal of Applied Science and Engineering*, *25*(12), 705–717. http://doi.org/10.53106/160792642022122307005

Sharma, S., & Malhotra, M. (2021). Web usage mining issues in big data: Challenges and opportunities. In *Impacts and challenges of cloud business intelligence* (pp. 102–112). IGI global. http://doi.org/10.4018/978-1-7998-5040-3.ch007

Singh, H., & Kaur, P. (2021). An effective clustering-based web page recommendation system. *SN Computer Science*, *2*(4), 736. http://doi.org/10.1007/s42979-021-00736-z

Soewito, B., & Johan, J. (2022). Website personalization using association rules mining. In *Conference on innovative technologies in intelligent systems and industrial applications* (pp. 689–698). http://doi.org/10.1007/978-3-031-29078-7_60

Sowbhagya, M. P., Yogish, H. K., & Raju, G. T. (2023). Perception based user profiles for web personalization. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(7S), 6986. http://doi.org/10.17762/ijritcc.v11i7s

Srivastava, A. K., & Srivastava, M. (2023). Irpdp_ht2: A scalable data pre-processing method in web usage mining using hadoop mapreduce. *Soft Computing*, *27*(2), 1001–1018. http://doi.org/10.1007/s00500-023-08019-w

Srivastava, M., Srivastava, A. K., & Garg, R. (2019). Data preprocessing techniques in web usage mining: A literature review. In *Proceedings of international conference on sustainable computing in science, technology and management (suscom)*. Amity University Rajasthan, Jaipur, India. http://doi.org/10.2139/ssrn.3352357

Srivastava, M., Srivastava, A. K., Garg, R., & Mishra, P. K. (2022). Performance evaluation of the mapreduce-based parallel data preprocessing algorithm in web usage mining with robot detection approaches. *IETE Technical Review*, *39*(4), 865–879. http://doi.org/10.1080/02564602.2021.1918584

Ting, I.-H., Tang, Y.-L., & Minetaki, K. (2024). Smart and adaptive website navigation recommendations based on reinforcement learning. *International Journal of Web and Grid Services*, *20*(3), 253–265. http://doi.org/10.1504/IJWGS.2024.139763

Wang, Z., & Li, J. (2023). Design of university archives business data push system based on big data mining technology. *International Journal of Advanced Computer Science and Applications*, *14*(11). http://doi.org/10.14569/ijacsa.2023.0141105

Waqas, M., Iram, M., Shahzad, S., Arshad, S., & Nawaz, T. (2018). Knowledge extraction using web usage mining. *ICST Transactions on Scalable Information Systems*, *5*(16), 154551. http://doi.org/10.4108/eai.13-4-2018.154551

Wasino, Lim, C., & Dewayani, E. (2023). User interface design of west java's intangible cultural heritage website using user-centered design. *International Journal of Application on Sciences, Technology and Engineering*, *1*(2), 421–432. http://doi.org/10.24912/ijaste.v1.i2.421-432

Win, T. N., & Lwin, N. K. Z. (2024). Analysis of customers' interest for web log clustering. In *Proceedings of the 21st ieee international conference on computer applications 2024 (icca 2024)* (pp. 219–224). http://doi.org/10.1109/ICCA62361.2024.10533033

Xing-hai, W. (2023). Reimagining website usability: A conceptual exploration of seo and ux design integration. *Journal of Digitainability, Realism & Mastery (DREAM)*, *2*(03), 60–66. http://doi.org/10.56982/dream.v2i03.99

Yau, N. Q., & Zainon, W. M. N. W. (2020). Understanding web traffic activities using web mining techniques. *International Journal of Engineering Technologies and Management Research*, *4*(9), 18–26. http://doi.org/10.29121/ijetmr.v4.i9.2017.96

Zagan, E., & Danubianu, M. (2023). Data lake architecture for storing and transforming web server access log files. *IEEE Access*, *11*, 40916–40929. http://doi.org/10.1109/access.2023.3270368

About the authors

Özkan Canay completed his PhD in the Department of Computer Engineering at Sakarya University, Türkiye, under the supervision of Prof. Dr. Ümit Kocabıçak. His research focuses on information systems, web usage mining, data analytics, user experience (UX) design and machine learning.

Ümit Koçabıçak served as a faculty member in the Department of Computer Engineering at Sakarya University, Türkiye, for many years. He is currently the president of the Turkish Higher Education Quality Council. His research interests include data analytics, machine learning, image processing, and computational modeling in engineering.