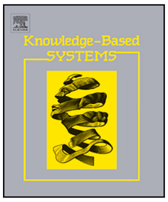
Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Predictive modeling and anomaly detection in large-scale web portals through the CAWAL framework

Özkan Canay [a,b,*], Ümit Kocabıçak [c,d]

[a] *Sakarya University of Applied Sciences, Vocational School of Sakarya, Dept. of Computer Tech., 54290, Sakarya, Turkiye*
[b] *Sakarya University, Institute of Natural Sciences, Dept. of Computer and IT Engineering, 54050, Sakarya, Turkiye*
[c] *Turkish Higher Education Quality Council, 06800, Ankara, Turkiye*
[d] *Sakarya University, Faculty of Computer and IT Engineering, Dept. of Computer Eng., 54050, Sakarya, Turkiye*

ARTICLE INFO

Keywords:
Web usage mining (WUM)
CAWAL
User behavior prediction
Anomaly detection
Machine learning

ABSTRACT

This study presents an approach that uses session and page view data collected through the CAWAL framework, enriched through specialized processes, for advanced predictive modeling and anomaly detection in web usage mining (WUM) applications. Traditional WUM methods often rely on web server logs, which limit data diversity and quality. Integrating application logs with web analytics, the CAWAL framework creates comprehensive session and page view datasets, providing a more detailed view of user interactions and effectively addressing these limitations. This integration enhances data diversity and quality while eliminating the preprocessing stage required in conventional WUM, leading to greater process efficiency. The enriched datasets, created by cross-integrating session and page view data, were applied to advanced machine learning models, such as Gradient Boosting and Random Forest, which are known for their effectiveness in capturing complex patterns and modeling non-linear relationships. These models achieved over 92% accuracy in predicting user behavior and significantly improved anomaly detection capabilities. The results show that this approach offers detailed insights into user behavior and system performance metrics, making it a reliable solution for improving large-scale web portals' efficiency, reliability, and scalability.

1. Introduction

Web usage mining (WUM) is the process of analyzing user interactions on websites to extract meaningful and valuable insights from their behavior. Web logs play a critical role by providing essential data such as navigation paths, pages visited, and interaction durations, enabling the analysis of user behaviors on websites [1]. Accurate analysis of user interactions is crucial for strategic decision-making, especially in fields like e-commerce, online education, and security [2]. However, the growing volume of data and the increasing complexity of user interactions challenge the capacity of traditional methods to process large datasets efficiently. Hence, integrating machine learning and data mining techniques into WUM offers significant opportunities, particularly in predictive modeling and anomaly detection, while also introducing new challenges [3].

Data preprocessing, one of the most critical stages in WUM, involves cleaning and organizing weblogs to extract meaningful information. However, traditional data preprocessing methods are time-consuming and complex, especially for large datasets [4]. For example, the use of social network analysis and frequent pattern mining to discover valuable information from extensive web data was proposed [5], while fuzzy techniques and clustering were focused on to understand user behavior in large datasets [6,7]. Despite these advancements, the need for more comprehensive and automated data processing techniques is growing.

Predictive modeling, a widely used WUM application, is a crucial method for forecasting future user activities on websites. In recent years, a study successfully applied Long Short-Term Memory (LSTM) networks to predict e-commerce users' shopping intentions with high accuracy [8]. Similarly, another recent study achieved high success in web page prediction using a chicken swarm optimization model based on neural networks [9]. These models contribute to strategic decision-making by predicting users' shopping tendencies and browsing habits. However, the accuracy of such predictive models is directly linked to the scope and richness of the datasets used. Due to their limited data coverage, web server logs often constrain these models' performance.

[*] Corresponding author at: Sakarya University of Applied Sciences, Vocational School of Sakarya, Dept. of Computer Tech., 54290, Sakarya, Turkiye.
E-mail address: canay@subu.edu.tr (Ö. Canay).
URL: https://canay.subu.edu.tr/ (Ö. Canay).

https://doi.org/10.1016/j.knosys.2024.112710
Received 29 September 2024; Received in revised form 24 October 2024; Accepted 3 November 2024
Available online 9 November 2024