Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 51, No. 1, 2025, pp. 37-51

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

Discussion on techniques of data cleaning, user identification, and session identification phases of web usage mining from 2000 to 2022

[1]Mohammed Ali Mohammed
College of Business Informatics
University of Information Technology and Communications
Baghdad, Iraq
Mohammed.ali@uoitc.edu.iq

[2]Hala Abdulsalam jasim
College of Science / Department of Remote Sensing and GIS
University of Baghdad
Baghdad, Iraq
hala.abd@sc.uobaghdad.edu.iq

[3]Ahmed Oday
College of Biomedical Informatics
University of Information Technology and Communications
Baghdad, Iraq
ahmed.oday@uoitc.edu.iq

ARTICLEINFO
Article History
Received: 31/01/2025
Accepted: 13/03/2025
Published: 31/5/2025

ABSTRACT

The data preprocessing step is an important step in web usage mining because of the nature of log data, which are heterogeneous, unstructured, and noisy. Given the scalability and efficiency of algorithms in pattern discovery, a preprocessing step must be applied. In this study, the sequential methodologies utilized in the preprocessing of data from web server logs, with an emphasis on sub-phases, such as session identification, user identification, and data cleansing, are comprehensively evaluated and meticulously examined.

Keywords:* *Web Usage Mining; Data Pre-processing Step; Access Log File.

1. INTRODUCTION

Web mining aims to elicit knowledge from various sources of web data, such as hyperlinks that employ patterns on different websites and web documents, by utilizing data mining strategies. Furthermore, web mining provides information about the process of uncovering patterns from the worldwide web. Web mining is regarded as an expansion of data mining and represents the integration of different strategies in research fields, such as knowledge discovery, artificial intelligence, informatics, statistics, and computational linguistics. The goal of web mining is to develop algorithms or techniques that increase the efficiency and convenience of data access [1][2].

As shown in Figure 1, depending on the part of the web that is being examined, the techniques used in web mining can be divided into three categories: web structure mining (WSM), web usage mining (WUM), and web content mining (WCM). WSM, the first type, uses graph theory to examine the relationships between webpages that are connected via either information or direct links. WUM, the second type, is the procedure that is intended to gain valuable insights into user activity by extracting patterns and information from server logs. WCM, the third type, involves extracting data from various sources on the worldwide web [3][4].

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

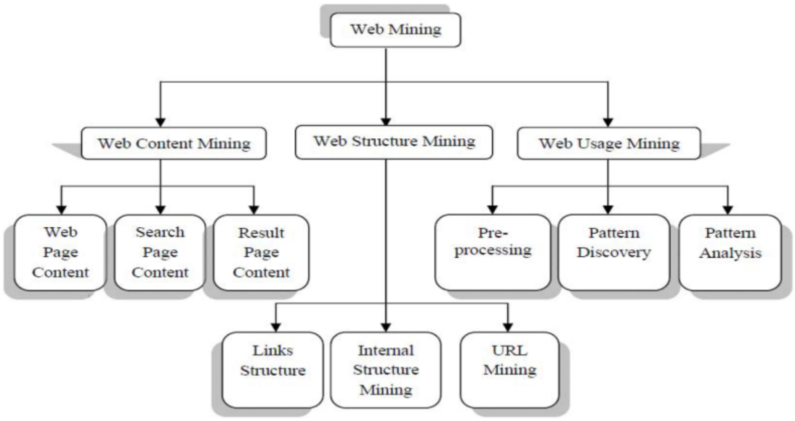

Fig. 1. Web Mining Category [3]

This type, also referred to as web log mining, is employed to examine how the users on the websites behave. This type explores the techniques that can be used to predict how users behave with the web during their interactions. This type also permits the collection of web access data for webpages. These data are collected automatically via the web server via access logs. WUM involves three processes. The first step, which is the preprocessing step, aims to eliminate noisy data and decrease the data size. Pattern discovery, which is the second step, uncovers web usage patterns by using a cleaned log file. The third step, Pattern Analysis, aims to extract more valuable data by analyzing patterns [3][5][6].

WUM has diverse applications, including enhancing website design, detecting intrusions, predicting user interests, identifying suspicious activities, optimizing e-commerce, analyzing the performance of the website, determining prime advertising locations, and analyzing social networks and traffic patterns [6][7][8].

Different research articles address WUM from various aspects, such as the tools used or the application, e.g., optimizing e-commerce or enhancing a website. The current work aims to address crucial inquiries regarding the latest studies of WUM.

This paper is organized as follows: Section 2 describes the log files, including their types and structures. Section 3 presents the datasets used in WUM. The preprocessing step is presented in section 4. A literature review for each study from 200 to 2022 is described in section 5. Finally, section 6 presents the conclusions of this study.

2. WEB LOG FILE STRUCTURE

The web log file records the different activities that the users perform on diverse websites; it has three primary sources: the proxy log file, the server log file, and the client log file [7].

Starting with the first source, the client log file captures the authentic actions of users. Nevertheless, obtaining the log file can be challenging, as it requires the cooperation of users [7][9]. Such log files can be stored within the client's browser window; this can be performed only after users download specific software in their browser window. Although these files exist in the client's browser window, only the web server has access to them [10][11]. The source-based approach has several disadvantages. For example, the design team needs to ensure successful deployment of the software and guide the end-users through the installation process. Thus, achieving compatibility with such diverse operating systems and web browsers can be challenging [12].

The second type, the proxy log file, contains a collection of cached pages or objects from different websites that have been accessed by different users; thus, the genuine picture of the behavior of the users is challenging to reveal [7][13]. A proxy server acts as a mediator between the user and the web server. Therefore, when the user receives a user request from the proxy server, the log file includes information about the proxy server instead of the original user. A separate log file is maintained by these web proxy servers to gather the user information [10][11]. The disadvantages of relying on the source are as follows: First, building a proxy server source can be challenging. Second, developed network programming, such as TCP/IP, is necessary for creating such a source. Finally, request interception is not extensive and covers only a limited number of requests.

When Web Quilt (web logging system) uses proxy logger implementation, its speed decreases because the proxy simulator must process each page request [12].

According to many researchers, the server log file is considered one of the most precise and accurate sources for WUM. The server log file consists of four commonly used types: referrer logs, access logs, agent logs, and

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

error logs [7][13]. The log file in the web server records the actions of users who access a website via their web browser. Similar to the abovementioned files, the contents of this file are crucial to ensure that the server used to collect users' personal information has a secure transfer mechanism [10][11]. Concerning this source, the disadvantages are as follows: First, these logs typically remain closed by the server owners owing to the sensitive and personal information they contain. Moreover, the logs do not keep track of previously accessed pages. Instead of being retrieved from web servers, cached pages are retrieved from the storage of browsers or proxy servers [12].

The referrer-related data are found in the referrer log. When the users navigate from one website to another, the server is expected to store the previous website and keep it in the referrer column. Google constantly utilizes the data of a log file [14][15].

The error log includes error messages such as "Error 404 File Not Found" or codes. This file is necessary for designers to enhance the performance of a website [14][15].

The agent log file contains records of the operating system used by website users, their browsers and what version they use. This information is valuable for individuals involved in website development and administrators, as it helps in creating websites that can work harmoniously with any web browser and different operating systems [14][15].

Given that all the hits, clicks, and accesses made by every website's users are documented in the access log file, it is regarded as a crucial log for web servers [10]. Thus, different types of data, such as user actions and preferences, how many times the users visit a page, and network traffic, can be collected from the log. Three main types of web server log file formats can be used to record the activities that are performed on websites by the users: extended log format (W3C), IIS log format (Microsoft), and common log file format (NCSA) [7][14][15]. Table 1 shows the descriptions of most attributes of the Common Log File Format [7][16].

Table I. Attributes of the Common Log File Format [7][16]

Attribute	Description
IP address/Hostname	User's IP address or host name
Rfcname	Provides users' authentication; a "-" character indicates that this field is empty
Logname	Provides the name that the user uses to log in; a "-" character indicates the absence of content in this field
Timestamp	Presents the current date and time when the user performs the request
HTTP_access_method	Describes the mode of request, which can be classified as GET, PUT, HEAD, or POST, among others
Requested_url	Presents the server's location on the requested page
HTTP_version	Returns the version of the HTTP protocol
HTTP_status_code	Provides the status of the response given by the server, e.g., HTTP status code 404 represents a "file not found on the server"
Page_size	indicates how many bytes are transferred from the server, e.g., the exact bytes of the requested web resource
Referrer_url	Provides the URL from which the requested page is accessed When a field value is lacking, it is represented by "-"
User_agent	Identifies users' operating systems and browsers' names and versions
Cookies	Refers to the information that is presented by a web server to the user regarding the details of a certain user

3. DATASETS

3.1 BABARAS HINDU UNIVERSITY DATASET

The log files are obtained from the web server of Banaras Hindu University. The logs are gathered from a blended log file with a specified input, which is "24/03/2014:00:00:00 to 30/03/2014:23:59:59," and then arranged chronically. Over the seven days, 6005814 is the total number of requests. Among 6005814 requests, the percentage of successful requests in the range of 200–299 is 80.82%, 300–399 is 8.41%, 400–499 is 10.75%, and above 500 is 0.001%. One of the most common errors that occurs is the 404 error, indicating that "file not found" accounts for 10.72% of all errors. Among the 6,005,814 requests, the majority (99.72%) contained the GET method. The other methods account for only 0.22%. At every stage, the sizes of the log files are consistently 1.28 GB [16][17].

3.2 TECHNICAL SCHOOL IN NOVI SAD DATASET

The log file was retrieved from the official website of the Advanced School of Technology in Novi Sad, which was created in 2009 in November; it can be accessed via the website http://www.vtsns.edu.rs/maja. The file adheres to the extended standard log format for analysis-related purposes. The raw log files comprise 12 attributes, including client IP, authenticated user, date and time, identification, request method, protocol version, URI-stem, status code, referrer, size in bytes, and user agent [18][19][20].

3.3 UNITED STATES EPA DATASET

Another type of data is a real-world dataset obtained from the United States EPA, which can be accessed via the website http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html, with a size of 4.4 MB. The format files and the standard log files are approximate mimics. Each operation performed by the user's browser and received by the EPA web server in Research Triangle Park, North Carolina, is represented as a line in this text file. An IP address, status code field, date/time field, HTTP request, and transfer size field are found for every record; it also presents quantitative details such as the period devoted to observing the webpage. This dataset was collected from 24 hours of HTTP requests. The dataset log for the EPA was observed from 23:53:25 EDT on August 29th, 1995, to 23:53:07 on August 30th, 1995. Among the 47748 requests, 46014 were GET requests, 1622 were POST requests, 107 were HEAD requests, and 6 were invalid requests [21][22].

3.4 MSNBC DATASET

The MSNBC log file contains a total of 17 subject categories and encompasses a wide range of topics, such as lifestyle, news, weather, health and medicine, music, and sports. This dataset incorporates 989,818 users, with each user averaging 5.7 pages. This dataset is available at http://archive.ics.uci.edu/ml/machine-learning-databases/msnbc-mld/msnbc.data.html [22].

3.5 NASA DATASET

The information in this dataset includes a one-month collection of server logs from the NASA.gov website. All the data were collected in August 1995. The data were compiled as supplementary data for an article in which the data were analyzed via SQL. For analysis purposes, a total of MB log files from the NASA web server are used. A server log file from NASA containing files of seven days with 195 MB size. It can be accessed and downloaded via https://www.kaggle.com/datasets/souhagaa/nasa-access-log-dataset-1995 [12][23].

4. DATA PREPROCESSING

Preprocessing is the essential first step, which is then preceded by other steps, such as the execution of data mining methods, such as classification and clustering, to extract valuable information and association rules. A variety of steps are entailed in data preprocessing, such as identifying users' data cleaning, identifying sessions, and completing paths [24][25].

As the preprocessing step is proficient, one can search for frequent patterns or specific rules within web data, even with limited time. It helps to examine web log data and differentiate their characteristics. The preprocessing phase and the development of an appropriate algorithm are necessary after the data have been fetched [24][25].

5. LITERATURE SURVEY

Ref. [26] (2000) utilized the cleaning parameter for multimedia files (image files) and robots' requests to combine W3C by checking recurring requests from a particular host for a particular URL and analyzing the time difference between two inquiries. Robot detection approaches are computationally expensive because they evaluate hosts whose referent URLs are all empty and invalid, and other cleaning criteria are ignored.

Ref. [27] (2001) reported that heuristics are typically based on referrers and requests that are considered to be part of a session. Thus, the referrer is classified as the undefined referrer whether it has already been obtained in the session or if the time interval between two consecutive requests is less than the delay. Otherwise, a new session is created, from which several few sessions are generated.

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Ref. [28] (2002) reported that the cleaning parameters are text files, admin files, multimedia files, script files, and flash animation files. Typically, websites that do not have links to other pages and employ frames can benefit from filtering, and no robot-detecting method is applied.

Ref. [29] (2002) explored the registration process where users log onto a website. The level of privacy is considered to be medium. The most reliable method of user identification is one that relies solely on user cooperation, and the least reliable method is considered. Many individuals tend to avoid registration websites.

Ref. [30] (2003) utilized W3C and various cleaning parameters to address issues such as multimedia files, script files, and style sheet files. The authors also addressed problems related to corrupt requests, inappropriate access methods, and robot requests. Data can be lost during the cleaning process when requests are made with improper access methods. Consequently, overly strict conditions are used for the request parameters. Thus, considering various HTTP status codes is crucial.

Ref. [31] (2003) utilized the IP address (reactive), which assumes that each unique IP address represents a unique user and that privacy concerns are minimal. This method has one advantage, which is always available, and one disadvantage, which is that it does not work with proxy servers.

Ref. [32] (2004) utilized a time-oriented heuristic that involved session time and page stay with a dynamic threshold. An adynamic threshold calculates the duration that a person consumes on a website in a single visit. The enhanced session data and dynamic threshold limit the duration of time spent on a page on the Internet.

Ref. [33] (2004) utilized the extended common logit function (ECLF) for performing cleaning parameters such as multimedia files, script files, and robot requests utilizing robots.txt, a compilation of user agents for popular robots, and the technique of optimizing browsing speed. The database of robots is not up to date, so additional data sources are needed to detect them, such as a list of commonly known robot user agents. User session time, which is unregistered in the web server log, is required by the browsing speed approach. Thus, unsuccessful requests are ignored.

Ref. [34] (2006) utilized the IP address finite user's inactive time (reactive) type, which limits the algorithm's stages to only users who are active, and the privacy concerns can be kept to a minimum. The method has favorable aspects, such as rapid algorithms and a Done-in-time complexity of O(n), but it has some drawbacks; that is, it also fails in the case of a proxy server.

Ref. [35] (2007) utilized the ECLF and various cleaning parameters, such as script files, irrelated access methods (except GET), multimedia files, robot requests (using the browsing speed method and robot.txt), and request outcomes. Owing to unsuccessful requests and inappropriate access methods, the information was lost during the cleaning process. This phenomenon can also occur when request parameters have conditions that are extremely strict.

Ref. [36] (2007) utilized the ECLF and a cleaning variable to analyze data, such as multimedia files (specifically image files), irrelevant access methods (excluding HTTP and POST), unsuccessful requests (excluding 200 and 299), robot queries (by referring to a list of user agents associated with known robots), and corrupt requests. Only relevant files are considered for deletion. Compiling a comprehensive updated list of all known robot user agents is a challenging task.

Ref. [37] (2008) utilized heuristic types (i.e., integer programming) and log registers to categorize requests coming from the same IP address. Enhanced session quality has been achieved, although it can be computationally expensive and might experience issues with browser caching or proxy caching.

Ref. [38] (2009) combined the IP address and user agent reactively, presuming that each individual user is identified by a unique IP address and user agent, which indicates low privacy concerns. This method has one disadvantage with respect to handling larger datasets. However, this approach has two advantages: it is always available, and it is capable of managing proxy server problems.

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Ref. [39] (2009), who applied back button surfing behavior, improved upon previous work by heuristically utilizing integer programming. Here, the session quality has been enhanced, although it is still significantly computationally demanding.

Ref. [40] (2009) adopted a two-way hashed table and employed a hash table as a heuristic type to trace the users' navigational history successfully, which can be challenging to obtain.

Ref. [41] (2010) relied on a structure-oriented approach where the access threshold of each page is defined according to its associated value. In contrast to fixed threshold approaches, the generated sessions exhibit a greater resemblance to genuine sessions. In contrast, the sequential approach proves to be ineffective when dealing with more giant datasets.

Ref. [42] (2011) used the ECLF along with various cleaning parameters, including multimedia files, unsuccessful requests (except 200), style sheet files, corrupt requests, reduction of log fields (IP address and requested URL are selected for later stages solely) and inappropriate access methods (except GET). Applying stringent requirements regarding the settings for reducing log fields, employing requests for inappropriate access approaches, and using unsuccessful requests can result in a loss of data throughout the cleaning phase.

Ref. [43] (2012) applied a time-oriented heuristic, especially session time and page stay, alongside a specific heuristic type, which is a navigation-oriented heuristic. A two-phased session is constructed, with the initial one employing session time and page stay time, whereas the latter follows the referrer rule. This scheme approach is an improvement over previously developed heuristics that emphasize time and navigation.

Ref. [44] (2012) used W3C and cleaning parameters, such as various types of requests (unsuccessful requests except 200–299), multimedia files, unsuccessful requests, corrupt requests, inappropriate access techniques (excluding GET), and robot requests, to handle their analysis. Given that the request parameters impose excessive constraints, information loss occurs throughout the cleaning process. Additional methods for robot detection are necessary to identify robots that fail to comply with the robot.txt criteria.

Ref. [45] (2013) used W3C and cleaning parameters, such as (multimedia files, script files, or unsuccessful requests except 200–200, and style sheet files), and the log data were reduced despite the need for any robot detection methods.

Ref. [46] (2015) used the ECLF and various cleaning parameters, including multimedia files (flash animation, audio, and picture files) and inappropriate access techniques. Excluding GET and POST, style sheet files, unsuccessful requests, script files, and corrupt requests were employed without engaging any methods to detect robots.

Ref. [47] (2015) examined one of the essential data preprocessing components: user session identification. Two methods were employed to identify users and sessions: "session time threshold (STT)" and "cookie identification." This study offers an in-depth examination of the methods mentioned above and their efficacy in identifying users and sessions.

Ref. [48] (2015) employed a fuzzy clustering framework, which depends on a mountain density function (MDF), to detect clusters of user sessions in web log data. This framework comprises various essential steps, such as preprocessing weblogs, using MDF to identify fuzzy user session clusters, and verifying the clusters in question. The clustering of user sessions is performed via fuzzy c-medoids and fuzzy c-means.

Ref. [49] (2015) employed sequential and content information and soft clustering, to enhance suggestion generation. This enhancement can be achieved via experimental examinations on various datasets, namely, the MSNBC benchmark dataset, the CTI dataset, and a simulated dataset. The approach and first-order Markov model and an arbitrary prediction method were compared.

Ref. [50] (2015) conducted a study that revolved around preprocessing and analysis of certain web data, which is "Dr. T.M.A. PAI polytechnic website," and discovery. A mixed model that combines neuro-fuzzy techniques is applied to obtain knowledge from website logs.

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Ref. [51] (2015) specifically studied the preprocessing steps of data fusion, data extraction, and data cleaning. An algorithm was presented by the researcher for gathering data; it precisely retrieves log data on the basis of the examination of the duration; it also efficiently and chronologically arranges the log entries, easing the prediction of sequences of the browsing of users. Then, the items extracted from an actual web server log are handled by the data cleaning algorithm. During the process of data cleaning, unnecessary files, irrelevant HTTP methods, and incorrect HTTP status codes are discovered and dismissed. Experiments have shown that the raw log data are reduced by approximately 80%, emphasizing the significance of the initial phases of data preprocessing.

Ref. [52] (2016) introduced a recursive method for clustering that incorporates density-based spatial clustering of applications with noise (DBSCAN) and expectation maximization (EM) algorithms to improve the reliability of clustering web-user sessions. As two types of patterns (frequent and sequential pattern mining techniques) are integrated, the detection of distinctive user access patterns is enhanced.

Ref. [53] (2016) presented a methodology that employs techniques for learning in an Internet-based multiagent application to reveal hidden patterns in the links that the users visit. This method uses a combination of different learning techniques and encourages collaboration among agents to identify patterns that organize the profiles of users on a selected website. The previously mentioned profiles are used to offer customized suggestions for the links and categories that interest the users.

Ref. [54] (2016) introduced a new method of analysis of clickstream data, that is, by employing a recommendation system and automatic web usage data extraction that employs K-nearest neighbor (KNN) classification, was investigated. This enables the delivery of personalized appropriate data to users on an RSS reader website.

Ref. [55] (2016) introduced D-ForenRIA, a distributed tool designed specifically for rich Internet application (RIA) session reconstruction. This tool shows the information pertaining to the DOM elements and user inputs involved and provides a comprehensive overview of the actions taken by users. By utilizing many browsers simultaneously, the system demonstrates adaptability for practical uses.

A semantically time-referrer-based strategy was used in [56] (2017) by the author to build the session. This strategy includes semantic data with reference and time-oriented heuristics. With respect to alternatives that lack semantic understanding, the sessions generated are more similar to actual sessions. However, when dealing with more extensive datasets, the sequential strategy is ineffective.

Ref. [57] (2017) investigated clickstream data analysis, with a specific focus on server log files as a valuable source of information about users' browsing habits. Such logs possess a vast amount of data, such as the constative users' clicks, the sequence of their navigation, and the time spent on each page. It delves into different mining techniques used for obtaining valuable insights from similar data. It investigates the wide range of applications that employ this type of analysis for obtaining valuable data.

Ref. [58] (2017) developed a method for extracting data and making predictions about the actions of users on e-commerce platforms. A web mining process is employed; it combines data gathered from the website's construction, extraction of semantic data, and examination of access data of the user. The suggested approach seeks to increase the precision of behavior prediction by considering different components of the website.

Ref. [59] (2018) comprehensively analyzed the structure of weblog data to shed light on the weaknesses of applying session reconstruction methods, such as session quality, pattern discovery, and reflected navigation. The methods used in the analysis of this study are agent-centric. However, the study proves that they are deemed insufficient for the specific situation. The suggested approach emphasizes the analysis of clickstreams, which usually occur between linked pages, taking into account the actual website topology. The quality of pattern discovery reflected by navigation and session quality is substantially enhanced by this approach in comparison with the agent-centric method.

Ref. [60] (2020) concentrated on the problem of gathering information from frequent user sessions in web log analysis. Frequent user session miner, which is an online algorithm, was developed to detect and retrieve sessions via frequent users in dynamic web data.

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Ref. [61] (2022) presented a method that uses reinforcement learning and probabilistic model-checking techniques to automate the generation of behavioral models. This method can be accomplished through two primary steps. First, on the basis of users' interactions, a set of probabilistic Markov models is generated dynamically. Second, the values of the reward to the state of the model are integrated. By using probabilistic model checking, this approach assesses the altering characteristics of the interaction pattern compared with the inferred behavioral models.

Ref. [62] (2024) proposed a new approach to determine request devices (computers or mobile devices) from platform information and operating system names. Additionally, the user sessions and complete paths were determined after the referrer URL was utilized. Finally, secure connections were proposed by determining robot URLs (e.g., search engines).
Table 2 presents the techniques and approaches, main focus, advantages, and disadvantages of all the articles mentioned in the literature review section.

TABLE II: Comparison of Techniques and Approaches

Reference	Technique/Approach	Main Focus	Advantages	Disadvantages
[26] Berendt & Spiliopoulou (2000)	Navigation behavior analysis	Integrating multiple information systems for analyzing user navigation on websites	- Comprehensive approach to web navigation - Can handle complex user navigation patterns	- Complexity in integrating multiple systems - May require substantial computational resources
[27] Berendt et al. (2001)	Sessionization techniques	Measuring the accuracy of sessionization methods in web usage analysis	- Provides quantitative evaluation of sessionizers - Focus on the accuracy of the user session reconstruction	- Limited to sessionization and may ignore broad WUM tasks - Dependent on specific sessionization methods tested
[28] Pabarskaite (2002)	Data cleaning and end-user interpretability	Cleaning web logs and improving interpretability of data for end-users	- Focuses on improving the usability of web log data for users - Enhances data quality via advanced cleaning techniques	- Potentially needs complex preprocessing steps - May reduce the data size, leading to the loss of some important patterns
[29] Anderson (2002)	Machine learning for web personalization	Personalized web experiences via machine learning	- Tailored user experience on the basis of data-driven insights - Uses machine learning for personalization	- Requires large datasets for effective personalization - Complexity in implementing machine learning models
[30] Yuan et al. (2003)	Data preprocessing algorithm	Preprocessing web log data for analysis	- Focuses on cleaning and transforming raw log data - Improves the quality of the data for mining tasks	- Preprocessing can be computationally expensive - Overcleaning may lead to the loss of valuable information
[31] Gery & Haddad (2003)	Predicting next user requests	Evaluating WUM approaches for request prediction	- Useful for enhancing the user experience by predicting the next steps - Helps in resource allocation and management	- Prediction accuracy may not always be reliable - May not handle unexpected user behavior effectively

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Reference	Technique/Approach	Main Focus	Advantages	Disadvantages
[32] Zhang & Ghorbani (2004)	Time-oriented heuristics	Reconstruction of user sessions via time-based heuristics	- More accurate sessionization with time-based heuristics - Focuses on the user navigation flow.	- Heuristics might not be generalizable across all types of web logs - Computationally expensive for large datasets
[33] Tanasa & Trousse (2004)	Data preprocessing for intersite web mining	Advanced preprocessing techniques for WUM across multiple sites	- Improves data quality across multiple websites - Can handle large-scale web log data	- May require extensive computational power - Complexity in implementing intersite mining techniques
[34] Khasawneh & Chan (2006)	Ontology-based web log mining	Active user-based and ontology-based data preprocessing	- Improves data understanding via ontology-based preprocessing - Offers have enhanced the interpretability of web log data	- Complexity in building and maintaining the ontology - May be unsuitable for all types of web logs
[35] Castellano et al. (2007)	Log data preprocessing tool	Tool for preprocessing web logs to mine browsing patterns	- Provides a comprehensive preprocessing tool - Facilitates pattern mining from web logs	- May not handle dynamic content well - Tool limitations may arise on the basis of data structure
[36] Liu & Kešelj (2007)	Combined mining of logs and web content	Combining web server logs with content analysis for user navigation classification	- Provides a more complete picture by combining logs and web content - Can predict user navigation patterns more effectively	- Requires access to both log and content data, which may not always be available - High complexity in terms of integration and processing
[37] Dell et al. (2008)	Integer programming for session reconstruction	Web user session reconstruction via integer programming	- Integer programming helps in precise session reconstruction - Effective for complex sessionization tasks	- Integer programming can be computationally intensive - May not scale well for very large datasets
[38] Suneetha & Krishnamoorthi (2009)	User behavior analysis via weblogs	Analyzing user behavior via web server access logs	- Helps in understanding user interactions and patterns - Improves web content by analyzing usage patterns	- May not fully capture dynamic or nonlinear user behaviors
[39] Dell et al. (2009)	Session reconstruction with back-button browsing	Reconstructing user sessions while accounting for back-button browsing	- Accounts for a critical aspect of web browsing behavior - Improves sessionization accuracy	- Increased complexity in handling user behaviors - Potentially slower session reconstruction process
[40] Arumugam & Suguna (2009)	User session sequence generation	Optimal algorithms for generating user session sequences	- Focuses on sequence generation for better session analysis	- Limited scalability for very large datasets

Reference	Technique/Approach	Main Focus	Advantages	Disadvantages
			- Enhances the understanding of the user navigation flow	- May require specific server-side log configurations
[41] Fang & Huang (2010)	Page-frame and page-threshold-based session identification	Session identification via frame pages and thresholds	- - Provides an effective method for sessionization - Easy to implement in most web environments	- May not work well for websites with complex navigation
[42] Aye (2011)	Web log cleaning for mining usage patterns	Cleaning web logs for improved pattern mining	- Improves new data quality and ensures accurate mining results - Helps in removing noise from the logs	- Overcleaning may result in the loss of useful data - May be time-consuming for large logs
[43] Dohare et al. (2012)	Novel WUM techniques	WUM with novel techniques for data preprocessing	- Introduces new methods to improve data quality - Enhances the accuracy of the mining results	- Techniques might be too novel for wider adoption - May require additional expertise to implement effectively
[44] Losarwar & Joshi (2012)	Data preprocessing in the WUM	Improving the preprocessing stage in the WUM	- Focuses on improving the data preprocessing phase - Ensures that web usage data are clean and reliable	- The preprocessing step may add significant overhead to the mining process
[45] Reddy et al. (2013)	Effective data preprocessing methods	Proposes effective methods for WUM data preprocessing	- Improves the accuracy and efficiency of the WUM - Reduces noise and unnecessary data during preprocessing	- Methods might be unsuitable for all types of web logs - Processing large datasets can be time-consuming
[46] Srivastava et al. (2015)	Data extraction and cleaning	Focus on the extraction and cleaning of web usage data	- Ensures high-quality input data for the WUM - Reduces the risk of incorrect analysis due to data noise	- Data cleaning can be a resource-intensive process - May not handle all types of web log anomalies effectively
[47] Aggarwal & Mangat (2015)	Application areas of the WUM	Overview of various applications of the WUM	- Provides a broad perspective on various applications - Highlights the impact of the WUM across industries	- Lacks a deep dive into specific mining techniques - Generalized conclusions may be inapplicable to all contexts
[48] Ansari et al. (2015)	Fuzzy-based web usage clustering	Mountain density-based fuzzy approach for web usage clustering	- Uses fuzzy logic to handle uncertain data - Helps in identifying patterns in user navigation behavior	- May be unsuitable for very large datasets - Fuzzy logic might introduce ambiguity in the clustering process
[49] Mishra et al. (2015)	Sequential information for web recommendation	Web recommendation system considering sequential data	- Enhances recommendation systems by incorporating sequential behavior	- Sequential data might not always be available

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Reference	Technique/Approach	Main Focus	Advantages	Disadvantages
			- Improves the accuracy of predictions	- Complex models may increase the computation time
[50] Shivaprasad et al. (2015)	Neuro-fuzzy hybrid model	Hybrid neuro-fuzzy model for the WUM	- Combines neural networks and fuzzy logic for enhanced accuracy - Offers better handling of uncertainties in data	- Complexity in implementing the hybrid model - Requires significant computational resources
[51] Srivastava et al. (2015)	Data extraction and cleaning	Focus on improving extraction and cleaning processes in WUM	- Ensures accurate data for analysis - Reduces noise in the data mining process	- Resource-heavy cleaning process - Unsuitable for dynamic web environments
[52] Udantha et al. (2016)	Combining EM and DBSCAN for user behavior modeling	User behavior modeling via the EM and DBSCAN clustering algorithms	- Combines two powerful clustering algorithms for better results - Effectively models complex user behavior	- May struggle with very large datasets - Requires parameter tuning for optimal performance
[53] Lotfy et al. (2016)	Multiagent- and learning-based WUM	Multiagent systems for learning-based WUM	- Uses a multiagent approach is used to enhance the mining process - Effectively handles dynamic environments	- Requires managing multiple agents, increasing complexity - May not scale efficiently in real-time applications
[54] Adeniyi et al. (2016)	K-nearest neighbor classification	Automated WUM and recommendation via KNN	- Easy to implement and understand - Fast classification and recommendation	- Performance may degrade with very large datasets - KNN is sensitive to noisy data
[55] Hooshmand et al. (2016)	D-ForenRIA tool for session reconstruction	Tool for reconstructing sessions from RIAs	- Specially designed for RIAs - Helps reconstruct sessions in complex environments	- Tools are specific to the RIA and may not be fully generalizable - May be complex to implement in some environments
[56] Kaur & Aggarwal (2017)	Semantic time-referrer-based sessionization	Improved sessionization with semantically enriched data	- Provides more accurate sessionization results - Improves the handling of dynamic content	- Complexity increases with semantic processing - Requires additional data enrichment steps
[57] Kapusta et al. (2017)	Session identification via STT and cookies	Session identification on the basis of timestamps and cookies	- Efficient session identification using time and cookies - Less computationally expensive than some other techniques	- May not handle users with disabled cookies effectively
[58] Dias & Ferreira (2017)	Automating static content and dynamic behavior extraction	Automating web content extraction for mining	- Automates the extraction of static and dynamic content - Enhances mining of user behavior	- Might miss some dynamic interactions - Needs real-time analysis

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Reference	Technique/Approach	Main Focus	Advantages	Disadvantages
[59] Ganibardi & Ali (2018)	Stream-centric approach for session reconstruction	Stream-centric approach for improving session reconstruction quality	- Efficiently handles session reconstruction in real time - Improves the overall quality of sessionization	- May not scale effectively with very large datasets
[60] Bei & Cai (2020)	Online session extraction for frequent users	Extracting sessions for frequent users	- Improves the extraction process for frequently active users - Optimizes for high-frequency users	- May ignore less active users, leading to incomplete analysis
[61] Ghaemmaghami et al. (2022)	Inferring user behavior models	Automatically inferring user behavior models for large-scale web apps	- Automatically adapts to large-scale environments - Reduces manual intervention	- Accuracy depends on the model quality - Needs extensive training data
[62] Mohammed et al. (2024)	Data collection and preprocessing in web mining	Analyzing data collection and preprocessing for the WUM	- Provides a detailed implementation and analysis - Helps refine preprocessing techniques	- Focuses more on implementation than novel techniques

6. CONCLUSION AND FUTURE WORK

Websites are crucial advertising tools for universities and other organizations in the international arena. Thus, by analyzing user interactions with a website, the quality of the website can be examined. WUM is used to evaluate user access to determine the quality of a website. Mining outcomes can be employed to improve the design of a website and increase user satisfaction, thereby benefiting a wide range of applications. In this study, different preprocessing steps, namely, data cleaning, user identification, session identification, and path completion, are discussed. Additionally, different datasets have been discussed. In addition, because more datasets are used, the Novi School dataset stands out as the most suitable choice for data mining. The preprocessing stage is an essential task in mining to ensure efficient pattern analysis. Achieving accurate mining results requires access to the user's session details (i.e., the mining is based on the user, and the user is based on the session). The survey examined several web usage strategies in preprocessing that have been offered by the scientific community.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Salman, R. H., Zaki, M., & Shiltag, N. A. (2020). A studying of web content mining tools. Al-Qadisiyah journal of pure science, 25(2), 1-16.

[2] Chandra, N., & Kumar, K. (2020). Designing Web Mining Technique for Customer Segmentation. Naresh Chandra. Journal of Engineering Research and Application, 10(2), 1-5.

[3] Gayatri, M., Satheesh, P., & Rao, R. R. (2018). Review of Current Trends in Web Usage Mining. International Journal of Engineering & Technology, 7(3.20), 690-694.

[4] Sellamy, K., Fakhri, Y., Boulaknadel, S., Moumen, A., Hafed, K., Jamil, H., & Lakhrissi, Y. (2018, April). Web mining techniques and applications: Literature review and a proposal approach to improve performance of employment for young graduate in Morocco. In 2018 international conference on intelligent systems and computer vision (ISCV) (pp. 1-5). IEEE.

[5] Mughal, M. J. H. (2018). Data mining: Web data mining techniques, tools, and algorithms: An overview. International Journal of Advanced Computer Science and Applications, 9(6).

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[6] Padigela, P. K., & Suguna, R. (2020). A survey on analysis of user behavior on the digital market by mining clickstream data. In Proceedings of the Third International Conference on Computational Intelligence and Informatics: ICCII 2018 (pp. 535-545). Springer Singapore.

[7] Nguyen, MT., Diep, TD., Hoang Vinh, T., Nakajima, T., Thoai, N. (2018). Analyzing and Visualizing Web Server Access Log File. In: Dang, T., Küng, J., Wagner, R., Thoai, N., Takizawa, M. (eds) Future Data and Security Engineering. FDSE 2018. Lecture Notes in Computer Science(), vol 11251. Springer, Cham. https://doi.org/10.1007/978-3-030-03192-3_27

[8] Yin Yin Htay, Myo Myo Khaing, Yamin. (2020). A STUDY ON WEB MINING FOR INFORMATION RETRIEVL TECHNIQUES. International Journal of Engineering Technology Research & Management. 4(3). 118-121.

[9] Murata, T., & Saito, K. (2006, December). Extracting users' interests from web log data. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06) (pp. 343-346). IEEE.

[10] Grace, L. K., Maheswari, V., & Nagamalai, D. (2011). Analysis of web logs and web user in web mining. arXiv preprint arXiv:1101.5668.

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

[12] Suneetha, K. R., & Krishnamoorthi, R. (2009). Identifying user behavior by analyzing web server access log file. IJCSNS International Journal of Computer Science and Network Security, 9(4), 327-332.

[13] Hussain, T., Asghar, S., & Masood, N. (2010, June). Web usage mining: A survey on pre-processing of web log file. In 2010 International Conference on Information and Emerging Technologies (pp. 1-6). IEEE.

[14] Preeti Rathi, Nipur Singh, Avanish Kumar. (2018). An Efficient Algorithm for Data Personalization to improve accuracy and performance of Web Usage Mining. International Conference on "Computing for Sustainable Global Development. Proceedings of the 12th INDIACom. 3652- 3657

[15] Al-Asdi, T. A., & Obaid, A. J. (2016). An efficient web usage mining algorithm based on log file data. Journal of Theoretical and Applied Information Technology, 92(2), 215.

[16] Srivastava, M., Garg, R., & Mishra, P. K. (2014). Pre-processing techniques in web usage mining: A survey. International Journal of Computer Applications, 97(18).

[17] Srivastava, M., Garg, R., & Mishra, P. K. (2015, March). 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).

[18] Hussain, W. A. G. A. (2017). Identifying of User Behavior from Server Log File. Iraqi Journal of Science, 1136-1148.

[19] Dimitrijevic, M., & Krunic, T. (2013). Association rules for improving website effectiveness: case analysis. Online Journal of Applied Knowledge Management (OJAKM), 1(2), 56-63.

[20] Mohd Khairudin, N., Mustapha, A., & Ahmad, M. H. (2014). Effect of temporal relationships in associative rule mining for web log data. The Scientific World Journal, 2014.

[21] Patil, U. M. (2020). User and Session Identification From Web Data Pre-processing. JAC: A Journal Of Composition Theory, 13(3), 376-385.

[22] Asadianfam, S., Kolivand, H., & Asadianfam, S. (2020). A new approach for web usage mining using case based reasoning. SN Applied Sciences, 2(7), 1251.

[23] Patel, D., & Bhatt, M. (2013). Indirect Positive and Negative Association Rules in Web Usage Mining. International Journal of Computer Applications, 69(24).

[24] Mehra, J., & Thakur, R. S. (2018). An effective method for web log pre-processing and page access frequency using web usage mining. Int. J. Appl. Eng. Res, 13(2), 1227-1232.

[25] Jayanti Mehra, R S Thakur. (2018). An Algorithm for user Identification for Web Usage Mining. International Journal of Innovations in Engineering and Technology (IJIET). 10(4). pp 110-115.

[26] Berendt, B., & Spiliopoulou, M. (2000). analysis of navigation behaviour in web sites integrating multiple information systems. The VLDB journal, 9, 56-75.

[27] Berendt, B., Mobasher, B., Spiliopoulou, M., & Wiltshire, J. (2001, April). Measuring the accuracy of sessionizers for web usage analysis. In Workshop on Web Mining at the First SIAM International Conference on Data Mining (pp. 7-14). Philadelphia PA: SIAM.

[28] Pabarskaite, Z. (2002, June). Implementing advanced cleaning and end-user interpretability technologies in web log mining. In ITI 2002. Proceedings of the 24th International Conference on Information Technology Interfaces (IEEE Cat. No. 02EX534) (pp. 109-113). IEEE.

[29] Anderson, C. R. (2002). A machine learning approach to web personalization. University of Washington.

DOI: https://doi.org/10.25195/ijci.v51i1.549
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[30] Yuan, F., Wang, L. J., & Yu, G. (2003, November). Study on data pre-processing algorithm in web log mining. In Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 03EX693) (Vol. 1, pp. 28-32). IEEE.

[31] Gery, M., & Haddad, H. (2003, November). Evaluation of web usage mining approaches for user's next request prediction. In Proceedings of the 5th ACM international workshop on Web information and data management (pp. 74-81).

[32] Zhang, J., & Ghorbani, A. A. (2004, May). The reconstruction of user sessions from a server log using improved time-oriented heuristics. In Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004. (pp. 315-322). IEEE.

[33] Tanasa, D., & Trousse, B. (2004). Advanced data pre-processing for intersites web usage mining. IEEE Intelligent Systems, 19(2), 59-65.

[34] Khasawneh, N., & Chan, C. C. (2006, December). Active user-based and ontology-based web log data pre-processing for web usage mining. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06) (pp. 325-328). IEEE.

[35] Castellano, G. I. O. V. A. N. N. A., Fanelli, A. M., & Torsello, M. A. (2007, February). LODAP: a log data preprocessor for mining web browsing patterns. In Proceedings of the 6th Conference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases (pp. 12-17).

[36] Liu, H., & Kešelj, V. (2007). Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data & Knowledge Engineering, 61(2), 304-330.

[37] Dell, R. F., Roman, P. E., & Velasquez, J. D. (2008, December). Web user session reconstruction using integer programming. In 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (Vol. 1, pp. 385-388). IEEE.

[38] Suneetha, K. R., & Krishnamoorthi, R. (2009). Identifying user behavior by analyzing web server access log file. IJCSNS International Journal of Computer Science and Network Security, 9(4), 327-332.

[39] Dell, R. F., Román, P. E., & Velásquez, J. D. (2009). Web user session reconstruction with back button browsing. In Knowledge-Based and Intelligent Information and Engineering Systems: 13th International Conference, KES 2009, Santiago, Chile, September 28-30, 2009, Proceedings, Part I 13 (pp. 326-332). Springer Berlin Heidelberg.

[40] Arumugam, G., & Suguna, S. (2009, June). Optimal algorithms for generation of user session sequences using server side web user logs. In 2009 International Conference on Network and Service Security (pp. 1-6). IEEE.

[41] Fang, Y., & Huang, Z. (2010, July). A session identification algorithm based on frame page and pagethreshold. In 2010 3rd International Conference on Computer Science and Information Technology (Vol. 6, pp. 645-647). IEEE.

[42] Aye, T. T. (2011, March). Web log cleaning for mining of web usage patterns. In 2011 3rd International Conference on Computer Research and Development (Vol. 2, pp. 490-494). IEEE.

[43] Dohare, M. P. S., Arya, P., & Bajpai, A. (2012). Novel web usage mining for web mining techniques. International Journal of Emerging Technology and Advanced Engineering, 2(1), 253-262.

[44] Losarwar, V., & Joshi, D. M. (2012, July). Data pre-processing in web usage mining. In International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July (pp. 15-16).

[45] Reddy, K. S., Reddy, M. K., & Sitaramulu, V. (2013, February). An effective data pre-processing method for Web Usage Mining. In 2013 International Conference on Information Communication and Embedded Systems (ICICES) (pp. 7-10). IEEE.

[46] Srivastava, M., Garg, R., & Mishra, P. K. (2015, March). 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).

[47] Aggarwal, S., & Mangat, V. (2015, February). Application areas of web usage mining. In 2015 Fifth International Conference on Advanced Computing & Communication Technologies (pp. 208-211). IEEE

[48] Ansari, Z., Sattar, S. A., Babu, A. V., & Azeem, M. F. (2015). Mountain density-based fuzzy approach for discovering web usage clusters from web log data. Fuzzy Sets and Systems, 279, 40-63.

[49] Mishra, R., Kumar, P., & Bhasker, B. (2015). A web recommendation system considering sequential information. Decision Support Systems, 75, 1-10.

[50] Shivaprasad, G., Reddy, N. S., Acharya, U. D., & Aithal, P. K. (2015). Neuro-fuzzy based hybrid model for web usage mining. Procedia Computer Science, 54, 327-334.

[51] Srivastava, M., Garg, R., & Mishra, P. K. (2015, March). 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).

[52] Udantha, M., Ranathunga, S., & Dias, G. (2016, April). Modelling website user behaviors by combining the EM and DBSCAN algorithms. In 2016 Moratuwa Engineering Research Conference (MERCon) (pp. 168-173). IEEE

[53] Lotfy, H. M., Khamis, S. M., & Aboghazalah, M. M. (2016). Multi-agents and learning: Implications for Webusage mining. Journal of advanced research, 7(2), 285-295.

[54] Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Applied Computing and Informatics, 12(1), 90-108

[55] Hooshmand, S., Faheem, M., Bochmann, G. V., Jourdan, G. V., Couturier, R., & Onut, I. V. (2016, October). D-ForenRIA: a distributed tool to reconstruct user sessions for rich internet applications. In Proceedings of the 26th Annual International Conference on Computer Science and Software Engineering (pp. 64-74).

[56] Kaur, N., & Aggarwal, H. (2017). A novel semantically-time-referrer based approach of web usage mining for improved sessionization in pre-processing of web log. International journal of advanced computer science and applications, 8(1).

[57] Kapusta, J., Munk, M., & Halvoník, D. (2017, November). Quality of Extracted Sequential Rules by Session Identification Using STT and Cookies. In 2017 European Conference on Electrical Engineering and Computer Science (EECS) (pp. 150-154). IEEE.

[58] Dias, J. P., & Ferreira, H. S. (2017). Automating the extraction of static content and dynamic behaviour from e-commerce websites. Procedia Computer Science, 109, 297-304.

[59] Ganibardi, A., & Ali, C. A. (2018, November). Weblog Data Structuration: A Stream-centric approach for improving session reconstruction quality. In Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services (pp. 263-271).

[60] Bei, Y., & Cai, Z. (2020, April). Online Extracting Sessions of Frequent Users. In 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA) (pp. 440-446). IEEE.

[61] Ghaemmaghami, S. S. S., Emam, S. S., & Miller, J. (2022). Automatically inferring user behavior models in large-scale web applications. Information and Software Technology, 141, 106704.

[62] Mohammed, M. A., Hamid, R. A., & AbdulHussein, R. R. (2024). Data Collection and Preprocessing in Web Usage Mining: Implementation and Analysis. Iraqi Journal for Computers and Informatics, 50(2), 54-74.