Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

Data Collection and Preprocessing in Web Usage Mining: Implementation and Analysis

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

[2] Rula A. Hamid
College of Business Informatics
University of Information Technology and Communications (UOITC)
Baghdad, Iraq
eng_rula_amjed@uoitc.edu.iq

[3] Reem Razzaq AbdulHussein
College of Business Informatics
University of Information Technology and Communications (UOITC)
Baghdad, Iraq
reem@uoitc.edu.iq

ARTICLEINFO
Article History
Received:07/04/2024
Accepted: 15/05/2024
Published: 16/11/2024

ABSTRACT
Data collection and data preprocessing are crucial stages in web usage mining, mainly because of the unstructured, diverse, and noisy nature of log data. During data collection, log file datasets are loaded and merged. Effective and comprehensive data preprocessing plays a vital role in ensuring the efficiency and scalability of algorithms used in the pattern discovery phase of web usage mining. This work aims to address these phases by introducing two innovative approaches. The first approach focuses on determining the device used for accessing the web, distinguishing between computers and mobile devices. The second approach aims to determine user sessions and complete paths by utilizing the referrer URL. The entire preprocessing pipeline has been implemented using the C# programming language, and the source code is available on GitHub at the following link: https://github.com/Mohammed91/Web-Usage-Mining.

Keywords: Access Log File; Data Collection Step; Data Preprocessing Step; Web Usage Mining;

1. INTRODUCTION

Web mining refers to the utilization of data mining methods for extracting valuable insights from web-related data, encompassing web documents, interconnections between documents, and website usage patterns. Its primary objective is to uncover patterns within the vast expanse of the World Wide Web (WWW). Web mining serves as an extension of data mining, incorporating diverse technologies from research domains such as artificial intelligence (AI), statistics, informatics, knowledge discovery, and computational linguistics. The ultimate aim of web mining is to develop algorithms and techniques that enhance the efficiency and convenience of accessing and utilizing web data[1][2]. Web Mining techniques are categorized (as shown in Fig. 1) [3] into three classes depend on which part to be mined, which are Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WSM)[4]. Web content mining involves the extraction of valuable information from diverse sources throughout the World Wide Web. This process focuses on discovering and analyzing data contained within web pages and other web-based resources. Web structure mining, on the other hand, involves analyzing the relationships and connections between web pages. By employing graph theory, it examines the interlinking structure of web pages, either through direct links or through the information web pages share. While Web usage mining aims to extract meaningful patterns and insights from server logs. These logs capture and record user activity on websites, providing valuable information about user behavior and preferences[5].

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

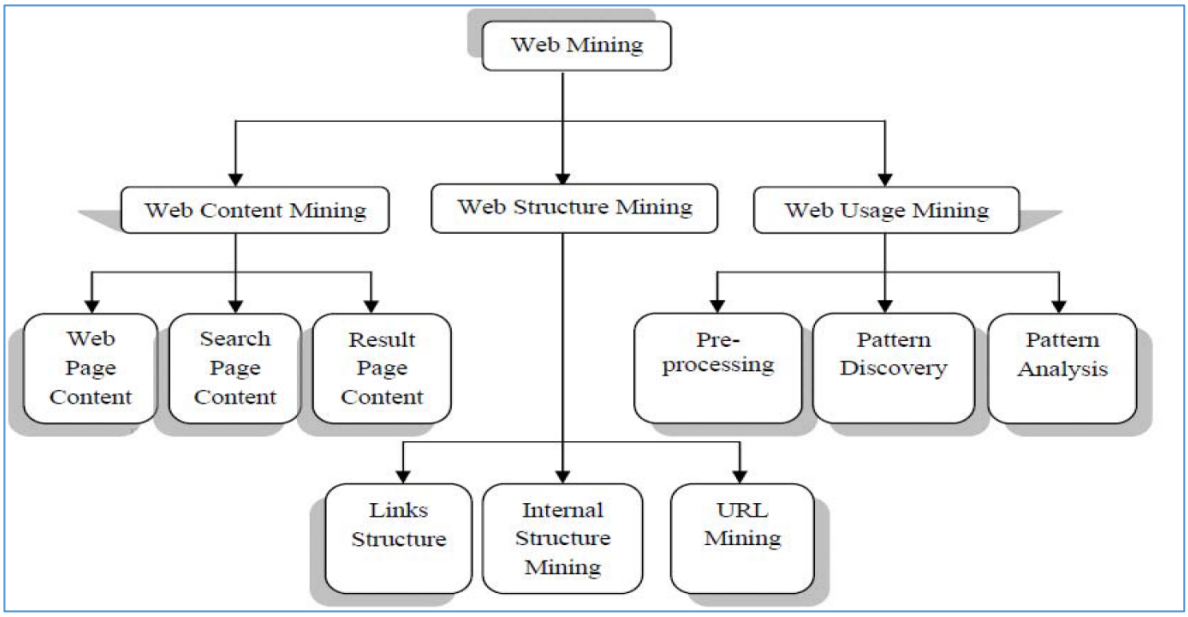

Fig. 1: Classification of Web Mining [3]

The Web usage mining is also known as Web Log mining, which is used to analyze the behavior of website users. This focuses on technique that can be used to predict the user behavior while user interacts with the web[5]. Web usage mining allows the collection of Web access information for Web pages, this information is often gathered automatically into access logs via the Web server. Web usage mining consists of three steps: Pre-processing: Removes noisy data and reduce the size of the data. Pattern discovery: Cleaned log file is used to discover web usage pattern. Pattern Analysis: Analyzing patterns in order to extract more useful information [6].

Web usage mining has its applications like improving websites design, intrusion detection, predicting user's interest, analyzing website performance, identifying suspicious activities, finding out primary places for advertising, and analyzing social network [7].

Data collection and preprocessing in web usage mining present several challenges because the data is often unstructured, diverse, and noisy, requiring careful extraction of meaningful information. Integrating and harmonizing datasets from different sources can be complex, and scalability becomes critical as the volume of web data increases. Other challenges related to Accurately identifying the device used for web access and reconstructing user sessions and paths due to the dynamic nature of web interactions[8]. Overcoming these challenges is vital to ensure reliable analysis and gain valuable insights from web data. Accordingly, this paper objectives to answer the important questions of emerging research of web using mining.

The research questions for this paper could be formulated as follows:

1-How can the device used for accessing the web be accurately determined, differentiating between computers and mobile devices, in the context of web usage mining?

2-How can user sessions and complete paths be effectively identified and reconstructed using the referrer URL in the preprocessing phase of web usage mining?

2. Literature Review

This section gives a comprehensive analysis of existing researches and studies related to web usage mining. It provides an overview of the current knowledge and improvements in the field by reviewing scholarly works, academic papers, and relevant publications.

The research query "Web usage mining" AND "user session" was applied to three reputable databases: IEEE, ACM, and Science Direct and the papers published between 2015 and 2023 were considered, resulting in a total of 7 papers from IEEE, 18 papers from Science Direct, and 13 papers from ACM(as shown in fig. 2). The abstract scan was done for the total (38) papers and the 15 papers were excluded from consideration as they do not align with the focus of our research, finally Full text reading was made for the remaining (23) papers and only (14) were included in

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

our review. This systematic approach was employed to analyze and categorize the collected papers based on their relevance, methodologies, and key contributions.

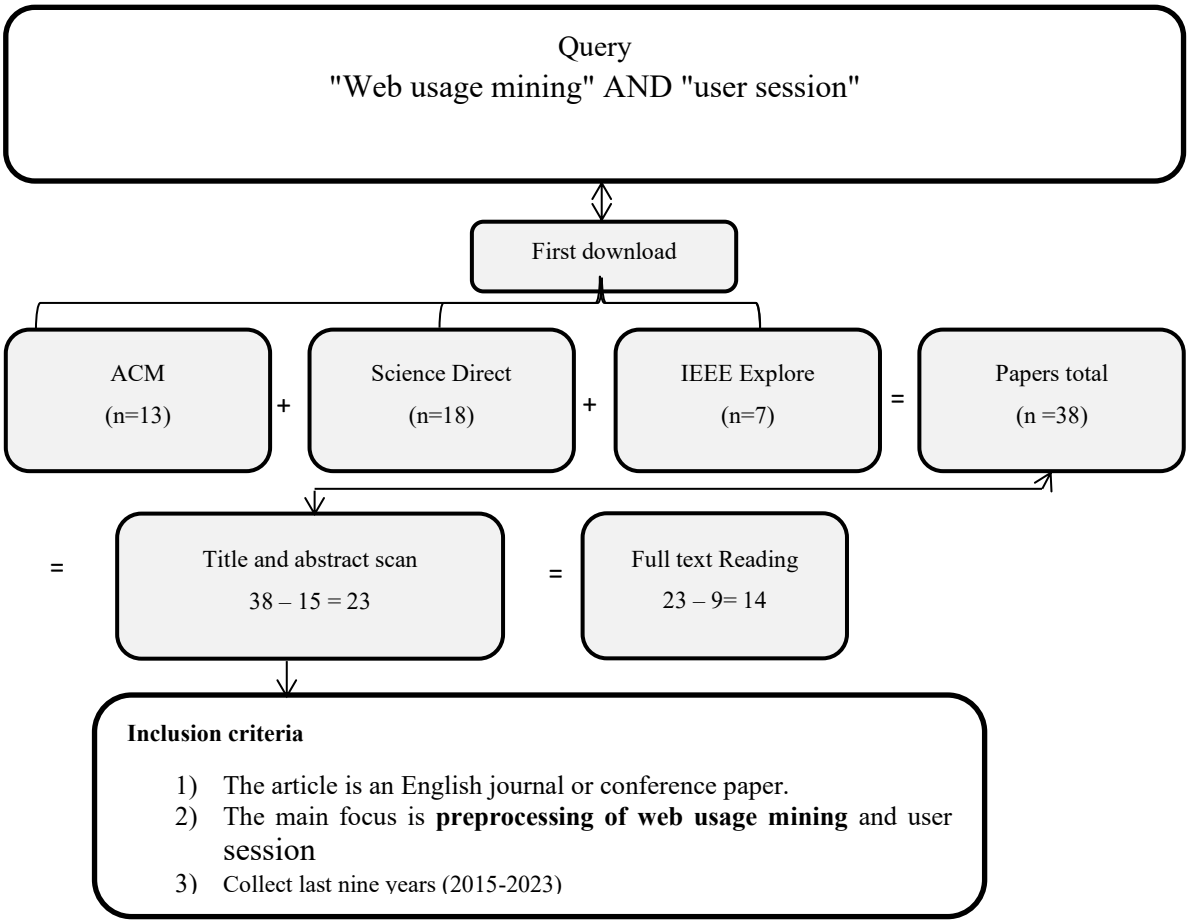

Fig. 2: Flowchart of Literature Review Query selection

The work in [9] aims to provide a comprehensive understanding of click stream data analysis, focusing on server log files as the primary source of valuable information about users' web usage. These logs capture a wealth of data, including the sequence of clicks made by users, the order in which they navigate through web pages, and the duration of their visits to each page. It explores various mining techniques employed to extract meaningful insights from such data and examines the diverse applications that leverage this analysis to derive valuable information. in[10] the authors explore the topic of user session identification, an essential step in data pre-processing and compares two methods which used for identifying users and sessions: Session Time Thresholds (STT) and cookie identification. The work provide a concise analysis of these approaches and their effectiveness in user and session identification. the work in [11] focuses on the problem of extracting sessions from frequent users in web log analysis. an online algorithm called FUSMiner was designed to identify and obtain sessions from frequent users in dynamic web data. the research in[12] was presented an iterative clustering technique that combines DBSCAN and EM algorithms, enhancing the accuracy of clustering web user sessions. By incorporating frequent pattern mining and sequential pattern mining, the identification of unique user access patterns is further improved. The authors in [13] introduced a methodology that utilizes learning techniques in a Web-based Multiagent application for uncovering hidden patterns in user's visited links. The approach combines unsupervised learning, reinforcement learning, and cooperation among agents to identify patterns that categorize the user's profiles within a sample website. The discovered profiles are utilized to provide personalized recommendations of interesting links and categories to the user. The work in [14]presented a fuzzy clustering framework based on a mountain density function (MDF) for identifying user session

clusters in web log data. The framework consists of several key steps, including web log preprocessing, utilizing MDF to discover fuzzy user session clusters, and validating these clusters. The clustering of user sessions is performed using the fuzzy c-means (FCM) and fuzzy c-medoids (FCMed) algorithms. in [15] the authors proposed an approach that automates the process of generating behavioral models by leveraging reinforcement learning and probabilistic model checking techniques. This is achieved through two main steps: 1) dynamically generating a set of probabilistic Markov

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

models based on users' interactions, and 2) incorporating reward values into the state of the model. the proposed approach estimate the interaction pattern's evolving properties against the inferred behavioral models using probabilistic model checking. The work in [16] utilized sequential and content information, along with soft clustering to improve recommendation generation through experimental evaluations on different datasets: the MSNBC benchmark dataset, a simulated dataset, and the CTI dataset. The approach comparison was made against the first order Markov model and a random prediction mode. The study [17]explored the application of automatic web usage data mining and a recommendation system that utilizes KNN classification to analyze click stream data, enabling the provision of personalized and relevant information to users on an RSS reader website. in [18] the work presented an approach for retrieving information and predicting user behavior on e-commerce websites. It utilizes a web mining process that incorporates data from the website's structure, semantic information extraction, and analysis of user access logs. The proposed approach aims to improve the accuracy of behavior prediction by considering multiple aspects of the website. The work in [19]focused on the pre-processing, discovery, and analysis of the Web Log Data from Dr. T.M.A.PAI polytechnic website. A hybrid model combining neuro-fuzzy techniques is used for Knowledge Discovery from the web logs. In [20], an in-depth examination of the structure of weblog data was conducted, highlighted the limitations of existing session reconstruction methods in terms of session quality, reflected navigation, and pattern discovery. The methods analyzed in the study are agent-centric, but they are found to be inadequate for the given context. The proposed method focused on clickstreams that occur between linked pages based on the actual website topology. This approach presents substantial benefits over the agent-centric structuration method, particularly in terms of improving session quality, reflected navigation, and pattern discovery quality.in [21] the authors introducing D-ForenRIA, a distributed tool designed for session reconstruction in Rich Internet Applications (RIAs). D-ForenRIA offers comprehensive insights into user actions, encompassing details about the involved DOM elements and user inputs provided. the system achieves scalability for real-world applications by utilizing multiple browsers simultaneously. The work in[22] concentrated on the preprocessing steps of data fusion, data extraction, and data cleaning. The authors were proposed an algorithm for data extraction that selectively retrieves log data based on an analysis of time duration. Additionally, the algorithm organizes log entries chronologically by date and time, facilitating the prediction of the user's browsing sequence. Subsequently, they employed a data cleaning algorithm to process the extracted real Web server log. During data cleaning, irrelevant files, irrelevant HTTP methods, and incorrect HTTP status codes are identified and removed. Through experimentation, they demonstrated the raw log data is reduced by approximately 80%, emphasizing the significance of the initial phases of data preprocessing. All comparing between these studies shown in table 1.

Table 1: compare between related work

Ref	method	dataset	work limitations
[9]	user/session with cookies and Session Time Thresholds methods	log file of university website contained 412360 records	The described experiment did not show a significant difference between the identification of users/sessions via cookies and the identification using the STT in terms of the quality of the extracted rules
[10]	Clustering Classification Association	-	Server log files-they don't mostly represent very accurate and consistent information of user click paths
[11]	FUSMiner	-	the FUSMiner efficiency and effectiveness may vary depending on the specific characteristics of the web environment and the quality of the input data, which could limit its generalizability across different contexts.
[12]	Spatial Clustering of Applications with Noise (DBSCAN(and Expectation Maximization (EM) algorithms	Log file	Limiting data exploration may result in incomplete understanding and potentially miss critical information..
[13]	Learning in Web-Based Education System	Weblog files are obtained from web servers' database	the proposed enhancements, introduce complexity and computational overhead,

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

			potentially affecting real-time performance or practical implementation in resource-constrained environments
[14]	mountain density function (MDF)- based fuzzy clustering framework	The dataset comprised web access logs extracted from the proxy servers of a university campus, spanning a week from June 1, 2008, to June 8, 2008. It encompassed a total of 723,263 web log requests.	the high dimensionality of user session data, which can make it difficult to accurately group sessions. Additionally, selecting suitable initial cluster centers is crucial for the clustering algorithms to work effectively,
[15]	reinforcement learning	requested URL submitted by a user	reinforcement learning may struggle with accurately representing certain nuances of user behavior, especially in scenarios where interactions are highly diverse or context-dependent.
[16]	similarity upper approximation an d singular value decomposition (SVD)	*MSNBC* benchmark dataset, simulated dataset and CTI dataset	reliance on historical user data to make predictions.
[17]	K-Nearest Neighbor	user behavior through his/her click stream data on the newly developed Really Simple Syndication (RSS) reader website,	-the system may face scalability issues when dealing with large volumes of web usage data, - the K-NN algorithm's computational complexity can increase significantly with the size of the dataset.
[18]	n all-in-one process	e-commerce website, we gathered data on the web structure, which includes: -A collection of 2,687 pages explored (nodes in the web graph). -A total of 361,344 functional links discovered (edges in the web graph).	crawling step introduces limitations in scalability and efficiency, as manual efforts may not be feasible for large-scale or rapidly changing websites.
[19]	Neuro-Fuzzy	Web Log Data of Dr. T.M.A.PAI polytechnic website, 5817 requests from 31st Dec. 2014 12:09:56 through 11:18:07 15th Jan. 2015, a total of 15 days.	The generated rules and patterns is difficult to interpret and understand, particularly for non-experts. This could limit the usability of the model for website administrators who may require clear and intuitive insights from the analysis.
[20]	stream-centric heuristic structuration	-	dependence on the availability and quality of web server logs. If the logs do not accurately capture user interactions or if they contain incomplete or noisy data, it could impact the effectiveness of the session construction process
[21]	D-ForenRIA(Rich Internet Applications)	sites with different technologies and from different domains.	It's unclear how to efficiently manage actions that don't produce HTTP traffic or requests that are cached by the browser.
[22]	an algorithm is proposed to extract data based on specified time intervals	-	lack of scalability

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

3. Web Log File Structure

When users surf various websites, their communications are captured and recorded in web log files. These log files serve as valuable sources of information about user behavior and website usage. There are three main sources (as shown in fig. 3) from which raw web log files are obtained: client log files, proxy log files, and server log files [23].

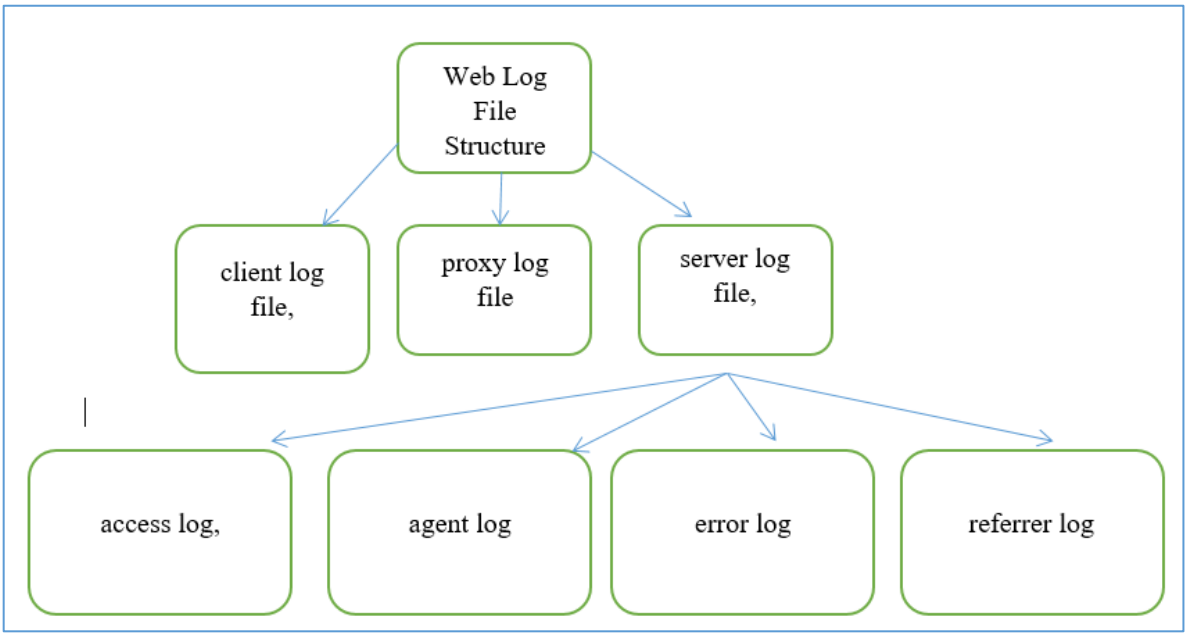

Fig. 3: Web Log File Structure

The client log file captures the most authentic user behavior, but acquiring client log file can be challenging due to the need for user collaboration. One potential solution is to store the log file within the client's browser window using specialized software that can be downloaded by the user[24]. While the log file resides in the client's browser the web server populates the log file with entries. In this way, the log file can still capture user behavior effectively, but the responsibility of logging entries is shifted to the web server[25]. This approach has drawbacks as it requires the design team to deploy specialized software and users to install it. Achieving compatibility with different operating systems and web browsers can be challenging also. These factors limit the ease of adoption and implementation of this approach [26].

The proxy log file contains cached pages or objects from multiple websites accessed by several users. However, extracting the accurate user behavior from these logs can be challenging[24]. Proxy servers act as intermediaries between clients and web servers so when a client sends a request through a proxy server to a web server, the log file entries will reflect the information of the proxy server rather than the original user. Proxy servers maintain separate log files to capture user information and activities[27]. This approach has certain drawbacks because constructing a proxy server is a complex task requires advanced network programming skills, particularly in TCP/IP. Additionally, the interception of requests is limited and may not cover all requests made by users. In cases where a proxy logger is implemented, such as in the Web Quilt logging system, the performance can decline as each page request needs to be processed by the proxy simulator, leading to potential performance issues [28].

The server log file in contrast, is the most reliable and accurate source for web usage mining. It provides valuable data for analyzing user behavior and website performance. The server log file typically consists of four main types: access log, agent log, error log, and referrer log[25].

The referrer log in the server log file contains information about the previous website from which a user accessed the current website. It captures the referrer and stores it in a designated column. This log file is commonly utilized by Google and serves various purposes. The error log in the server log file consists of error messages or codes, such as the commonly encountered "Error 404 File Not Found" message. It is particularly valuable for website designers as it provides insights into optimizing website links and resolving errors. The agent log file records details about the users' browsers, browser versions, and operating systems. This information is highly beneficial for website designers and

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

administrators as it helps them to develop websites that are compatible with different operating systems and web browsers. By analyzing the agent log file, they can ensure better website functionality and user experience across a wide range of platforms[29][30].

The access log file is the major log of web server which records all the clicks, hits and accesses made by any website user .There is a little useful information which can be extracted from the log, such as user behavior and interest, network traffic, the number of page views[31]. There are three main types of web server log file formats available to capture the activities of users on websites. They are Common Log File Format by (NCSA), Extended Log Format by (W3C) and IIS Log Format by (Microsoft) [32].Table 2 shows the description of the most common attributes of Log File Format [33].

Table 2: attribute descriptions of log file [33]

Feature name	Feature Description
IP address/Host name	Users' IP address or Host name.
Rfcname	Provides users' authentication. A "-"character shows that this field is empty.
Logname	Provides the login name of user. A "-"character shows that this field is empty.
Timestamp	Returns the date and time of users' request.
HTTP_access_method	Describes the mode of request. It can be GET, PUT, HEAD,POST etc.
Requested_url	Gives the path of requested page on server.
HTTP_version	Returns the version of HTTP protocol.
HTTP_status_code	Provides the status of response given by server e.g. HTTP status code 404 represents "file not found on server".
Page_size	Denotes the number of bytes transferred from the server i.e. the size of requested Web resource.
Referrer_url	Provides the URL from where requested page is accessed. When this field value is not present it is shown by "-".
User_agent	Identifies users' operating system and browsers' name and version
Cookies	It is a piece of information sent to user by Web server to identify the details of a particular user.

4. Technical School in novi sad dataset used

The dataset used in this study is derived from the official website of the Advanced School of Technology in Novi Sad, collected in November 2009[34]. The log file, which can be accessed at http://www.vtsns.edu.rs/, Fig. 4 show data sample of this dataset) .

The dataset has been chosen for analysis and research purposes to gain insights into user behavior and website usage patterns on the school's website. By studying this log file, valuable information can be extracted and analyzed to improve the website's performance and user experience. The raw log files consist of 9 attributes. Table **3** shows the names of dataset fields with their descriptions, datatypes and examples .

```
147.91.173.31 - - [16/Nov/2009:00:02:23 +0100] "GET / HTTP/1.0" 200 3669 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102 Firefox/3.5.5"
147.91.173.31 - - [16/Nov/2009:00:02:24 +0100] "GET /style.css HTTP/1.0" 200 6554 "http://www.vtsns.edu.rs/" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102 Firefox
147.91.173.31 - - [16/Nov/2009:00:02:24 +0100] "GET /img/vts.jpg HTTP/1.0" 200 4053 "http://www.vtsns.edu.rs/" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102 Firef
147.91.173.31 - - [16/Nov/2009:00:02:24 +0100] "GET /img/copyprint.jpg HTTP/1.0" 200 1662 "http://www.vtsns.edu.rs/" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102
147.91.173.31 - - [16/Nov/2009:00:02:24 +0100] "GET /img/novosti.jpg HTTP/1.0" 200 1161 "http://www.vtsns.edu.rs/" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102 F
147.91.173.31 - - [16/Nov/2009:00:02:24 +0100] "GET /img/head2.jpg HTTP/1.0" 200 74512 "http://www.vtsns.edu.rs/" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102 Fi
147.91.173.31 - - [16/Nov/2009:00:02:24 +0100] "GET /img/bg.jpg HTTP/1.0" 200 23084 "http://www.vtsns.edu.rs/style.css" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091
147.91.173.31 - - [16/Nov/2009:00:02:25 +0100] "GET /favicon.ico HTTP/1.0" 404 - "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102 Firefox/3.5.5"
147.91.173.31 - - [16/Nov/2009:00:02:28 +0100] "GET /konsultacije.php HTTP/1.0" 200 2822 "http://www.vtsns.edu.rs/" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102
147.91.173.31 - - [16/Nov/2009:00:02:28 +0100] "GET /img/head5.jpg HTTP/1.0" 200 23689 "http://www.vtsns.edu.rs/konsultacije.php" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) G
147.91.173.31 - - [16/Nov/2009:00:02:28 +0100] "GET /img/konsultacije.jpg HTTP/1.0" 200 2567 "http://www.vtsns.edu.rs/konsultacije.php" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.
147.91.173.31 - - [16/Nov/2009:00:02:31 +0100] "GET /konsultacije.doc HTTP/1.0" 200 117760 "http://www.vtsns.edu.rs/konsultacije.php" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.
147.91.173.31 - - [16/Nov/2009:00:02:51 +0100] "GET /vesti.php HTTP/1.0" 200 3367 "http://www.vtsns.edu.rs/konsultacije.php" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/
```

Fig. 4: Sample of Technical School in Novi Sad Dataset

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Table 3: fields descriptions of technical school in novi sad dataset

Filed Name	Description	Example
Remote Host	This field consists of the Internet IP address of the remote host making the request, such as "141.243.1.172". If the remote host name is available through a DNS lookup, this name is provided, such as "wpbfl2-45.gate.net."	147.91.173.31
Identification	This field is used to store identity information provided by the client only if the web server is performing an identity check.	-
Auth_user	This field is used to store the authenticated client user name, if it is required.	-
Date/Time	This field contains date and time of the request from the user's browser to the web server.	[16/Nov/2009:00:02:23 +0100]
HTTP Request	The HTTP request field consists of the information that the client's browser has requested from the web server. Contains: Request Method, URI, Header, Protocol.	"GET / HTTP/1.0"
Status Code	Not all browser requests succeed. The status code field provides a three-digit response from the web server to the client's browser, indicating the status of the request.	200
Transfer Volume (Bytes)	The transfer volume field indicates the size of the file (web page, graphics file, etc.), in bytes, sent by the web server to the client's browser.	3669
Referrer	The referrer field lists the URL of the previous site visited by the client, which linked to the current page.	-
User Agent	The user agent field provides information about the client's browser, the browser version, and the client's operating system.	"Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102 Firefox/3.5.5"

5. Data Collection

Data collection in web usage mining refers to the process of gathering relevant web access log information, which is performed from the server side .this is the initial step in web usage mining process where client log information is gathered from sources located on the server side, typically from the access log file [35].

In our implementation the initial step involves reading the unstructured log file from the dataset and displaying its content in a RichTextBox tool. However, a significant challenge arises due to the unstructured nature of the log file, making it difficult to process. Therefore, it becomes necessary to convert the raw data into a structured format. To achieve this, we adopt a straightforward approach by splitting each raw in the dataset into individual parts and storing the structured data in our MSSQL database. the database schema shown in fig. 5, The unprocessed data is stored in a table that preserves the raw data in its original form, without any cleaning or modifications.

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

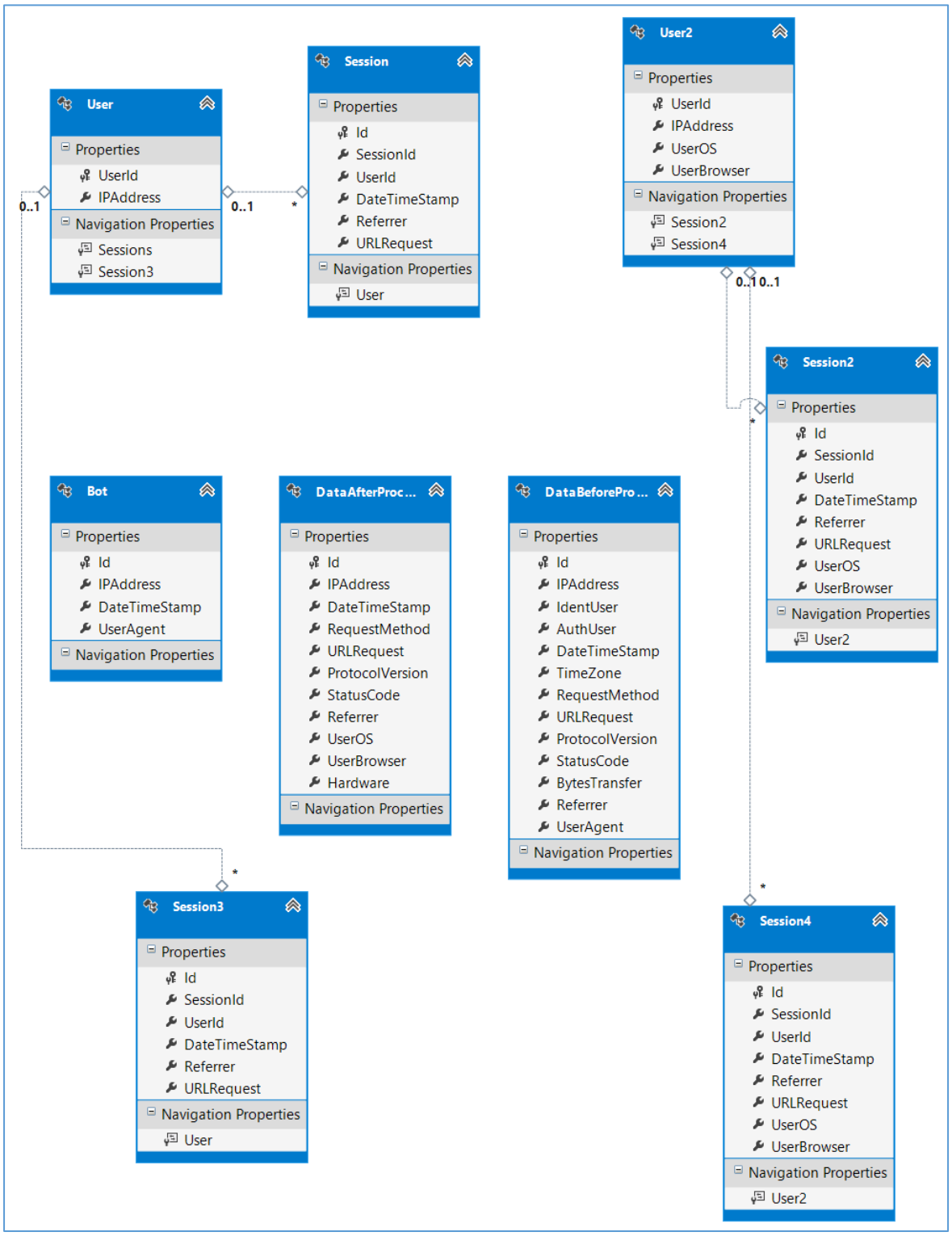

Fig. 5: Database Schema

This database table serves as a repository for the initial data collection stage. Once the data goes through the preprocessing step, it is stored in a separate table that holds the processed data. This database table contains the transformed and standardized information ready for further analysis or use in subsequent stages of the web usage mining process. The table called "Bot" is utilized to store the requests made by automated bots or robots. This allows for easy identification and analysis of robot-related activities in the web access logs. To track and analyze user behavior, two tables named "User" and "User2" are employed. These tables store user access information categorized

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

by IP address, operating system (OS), and web browser. This data can provide insights into the preferences and characteristics of different user segments. Furthermore, different types of sessions are recorded and stored in separate tables: Session, Session 2, Session 3, and Session 4. These tables capture session-specific data, allowing for analysis of various session characteristics and patterns within the web usage data. Finally, the result after convert raw data to structure data are shown in the fig. 6.

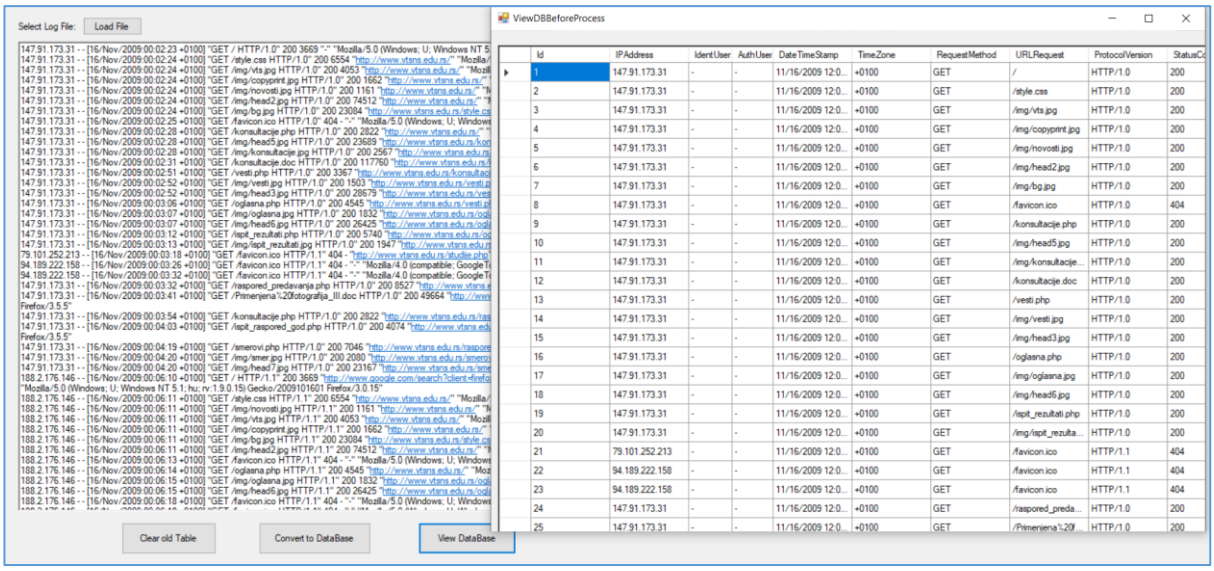

Fig. 6: The Result of Data Collection Step

6. Data Preprocessing

Data preprocessing is the step in which the data is cleaned, transformed, and prepared for analysis. It involves tasks like removing errors and inconsistencies, identifying individual users, grouping interactions into sessions, and completing missing information in user paths. After preprocessing, data mining methods like association rules, classification, and clustering are applied to uncover valuable insights and patterns within the prepared data[36].

Once the preprocessing of web log data is efficiently performed, it becomes easier to quickly search for frequent patterns or interesting rules within the data[37]. The preprocessing step involves analyzing the attributes of the web log data and extracting relevant information. After fetching the data, it is necessary to preprocess it to clean, transform, and organize the data, making it suitable for further analysis and exploration of patterns or rules of interest within the limited timeframe available[38]. Fig. 7 shows the Flowchart of pre-processing steps followed by this work and illustrates the flowchart for the preprocessing step in the web log data analysis. The purpose of this step is to prepare the raw data for further analysis by performing various operations and transformations as explained in the following steps.

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

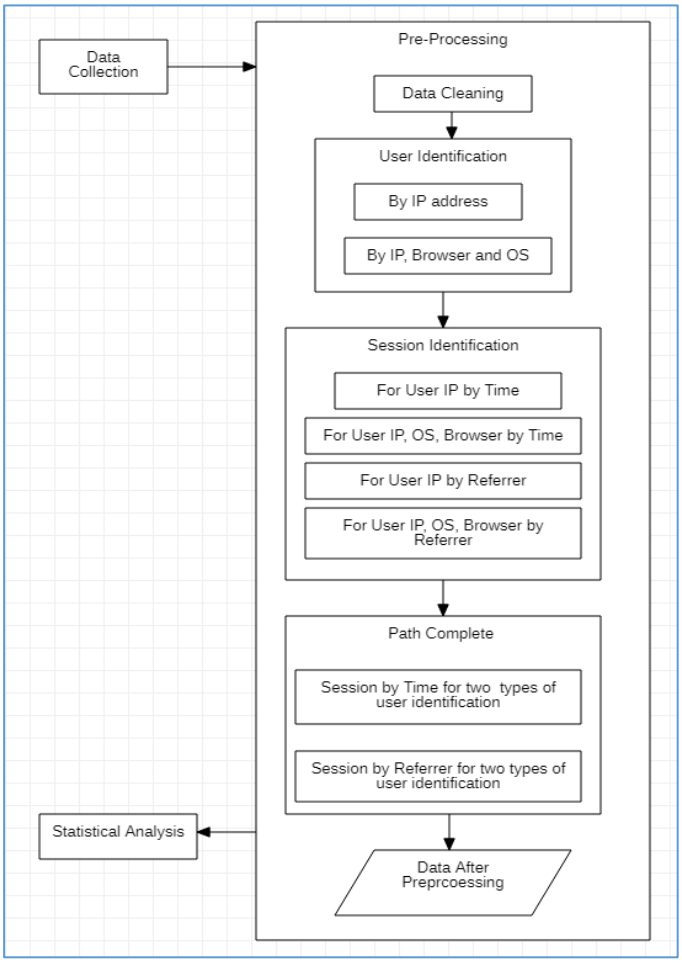

Fig. 7: Flowchart of pre-processing steps

6.1 Data Cleaning

Firstly the data was preprocessed by data cleaning .it is the process of removing unwanted or irrelevant data from web log files, such as non-relevant file types (e.g., GIF, JPEG, video, audio, CSS), and filtering out log entries with HTTP status codes outside the valid range (less than 200 or greater than 400) [39]. This step aims to eliminate noise and ensure that only accurate and meaningful records are retained. By identifying and correcting imprecise or inaccurate entries, data cleaning enhances the overall quality and reliability of the dataset, making it more suitable for subsequent analysis[40].

In our implement, when applying data cleaning steps ,the following data were removed from the database:

1) Remove columns: Remove the following columns from the database: IdentUser, TimeZone, AuthUser, and BytesTransfer.
2) Filter extensions: Iterate through each log entry and remove all entries that have file extensions other than web page extensions (e.g., HTML, PHP, ASP).
3) Filter status codes: Iterate through each log entry and remove all entries with status codes that are not within the range of 200-299 (indicating successful requests). Keeping only the log entries with status codes in this range.
4) Filter request methods: Iterate through each log entry and remove all entries with request methods other than GET and POST. Keeping only the log entries with these methods.
5) Split User Agent: Iterate through each log entry and split the user agent field to extract user OS and user browser information. a new approach was Proposed to determine the hardware type (computer or mobile) based on the

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

user OS by Identifying and categorizing user OS values as either computer OS (Windows, Linux, Mac) or mobile OS (Symbian, 2ME/MIDP).

6) Remove zero-byte records: Iterate through each log entry and remove all entries with a zero value in the bytes transferred field.

7) Filter out Robots: Identify and remove any Robot entries, such as those corresponding to a web crawler ("/robots.txt"), from the log data.

Web Robots[41], also known as web crawlers or spiders, are software programs search the internet for resources by clicking on links. Therefore, it is crucial to detect and remove them from log files. By eliminating web robots, the log data can be reduced, and the analysis can focus on genuine user access patterns. This helps in identifying and understanding the behavior of users, leading to more accurate insights and decision-making based on the log data. Removing web robots from log files improves data integrity and enhances the ability to discern genuine user activities. After implementing the above-mentioned data cleaning steps, the resulting data takes the form shown in Fig. 8.

Id	IPAddress	DateTimeStamp	RequestMethod	URLRequest	ProtocolVersion	StatusCode	Referrer	UserOS	UserBrowser	Hardware
1	147.91.173.31	11/16/2009 12:0...	GET	/	HTTP/1.0	200	-	Windows NT 5.1	Firefox/3.5.5	Computer
2	147.91.173.31	11/16/2009 12:0...	GET	/konsultacije.php	HTTP/1.0	200	http://www.vtsns...	Windows NT 5.1	Firefox/3.5.5	Computer
3	147.91.173.31	11/16/2009 12:0...	GET	/vesti.php	HTTP/1.0	200	http://www.vtsns...	Windows NT 5.1	Firefox/3.5.5	Computer
4	147.91.173.31	11/16/2009 12:0...	GET	/oglasna.php	HTTP/1.0	200	http://www.vtsns...	Windows NT 5.1	Firefox/3.5.5	Computer
5	147.91.173.31	11/16/2009 12:0...	GET	/ispit_rezultati.php	HTTP/1.0	200	http://www.vtsns...	Windows NT 5.1	Firefox/3.5.5	Computer
6	147.91.173.31	11/16/2009 12:0...	GET	/raspored_preda...	HTTP/1.0	200	http://www.vtsns...	Windows NT 5.1	Firefox/3.5.5	Computer
7	147.91.173.31	11/16/2009 12:0...	GET	/konsultacije.php	HTTP/1.0	200	http://www.vtsns...	Windows NT 5.1	Firefox/3.5.5	Computer
8	147.91.173.31	11/16/2009 12:0...	GET	/ispit_raspored_g...	HTTP/1.0	200	http://www.vtsns...	Windows NT 5.1	Firefox/3.5.5	Computer
9	147.91.173.31	11/16/2009 12:0...	GET	/smerovi.php	HTTP/1.0	200	http://www.vtsns...	Windows NT 5.1	Firefox/3.5.5	Computer
10	188.2.176.146	11/16/2009 12:0...	GET	/	HTTP/1.1	200	http://www.goog...	Windows NT 5.1	Firefox/3.0.15	Computer

Fig. 8: Data Cleaning Result

Table 4 on the other hand provides the statistical summary of the data after the application of the data cleaning. It reveals that the initial dataset contained 5,982 rows, while the dataset after the cleaning step consists of 1,606 rows. This indicates that a total of 4,376 rows were removed from the data during the cleaning process. These removed rows were considered unwanted due to factors such as file extensions like JPG, status codes of 400, and JavaScript (JS) file extensions. The data cleaning step effectively eliminated these unwanted rows, resulting in a more refined and focused dataset for further analysis.

Table 4: statistical values after data clean step

Key	Value
No. Zero Byte	1339
No. of Status Code 100	0
No. of Status Code 300	630
No. of Status Code 400	672
No. of Status Code 500	0
No. of PROPFIND	23
No. of OPTIONS	85
No. of HEAD	10
No. of jpg	2736
No. of gif	1
No. of png	45
No. of bmp	0
No. of CSS	345
No. of XML	1
No. of js	2
No. of ico	564
No. of doc	286
No. of ZIP	14
No. of rtf	44
No. of pdf	160
No. of Record Before	5982
No. of Record After	1606

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Finally, the result of Robot request values (11 robots unique) in the data set shown in Table 5. For each request will be comparing with the all requests in the /robots.txt file [42] if match, the request will be considering it as Robot request.

Table 5: robot request

IP Address	Time	User Agent
89.143.229.115	11/16/2009 1:08 AM	Mozilla/5.0 (compatible; Pogodak.co.yu/3.1)
83.167.62.168	11/16/2009 2:01 AM	Mozilla/5.0 (compatible; Exabot/3.0 (BiggerBetter); +http://www.exabot.com/go/robot)
67.195.112.25	11/16/2009 7:24 AM	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
67.218.116.164	11/16/2009 9:46 AM	Mozilla/5.0 (Twiceler-0.9 http://www.cuil.com/twiceler/robot.html)
67.202.15.176	11/16/2009 11:51 AM	archiver (+http://www.alexa.com/site/help/webmasters; crawler@alexa.com)
66.249.68.2	11/16/2009 12:38 PM	Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
66.235.124.17	11/16/2009 7:13 PM	Mozilla/5.0 (compatible; Ask Jeeves/Teoma; http://about.ask.com/en/docs/about/webmasters.shtml)
119.63.198.24	11/16/2009 7:33 PM	aiduspider+(+http://www.baidu.com/search/spider.htm)
66.249.68.50	11/16/2009 10:13 PM	Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
85.10.36.115	11/16/2009 10:35 PM	Mozilla/5.0 (compatible; Pogodak.co.yu/3.1)
220.181.7.54	11/16/2009 11:06 PM	aiduspider+(+http://www.baidu.com/search/spider.htm)

6.2 User Identification

The identification of a new user can be determined by examining the assigned IP address and the web agent they use. However, if both of these factors are identical, the new user can still be identified by analyzing the referral pages they used to search for information. The user's requests may not always be directed towards the intended web pages; sometimes they may refer to reference pages in order to reach their desired page[43]. The identification of the user is based on various factors such as the client IP address, user name, requested URL, date, time, and server IP address. The data is recorded in the log file format by the Internet Information Service (IIS)[44].

In our implementation, two distinct approaches were employed to identify users within the dataset. The first approach focuses on extracting unique IP addresses from the dataset. There are two ways to compute a number of users: firstly, the number of users (unique) according to only IP address is 297 users, and a number of users according to IP address, user operating system, and user browser is 307 users. The results of these ways are shown in Fig 9.

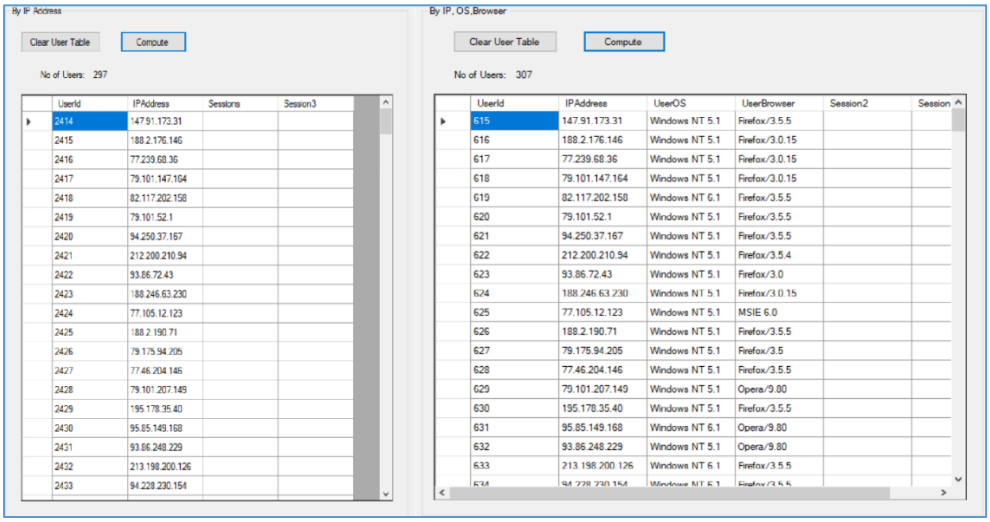

Fig 9: Results of user identification

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

6.3 Session Identification

From Session identification, the user session can be computed by determine the login time and logout time in the web access log file. It helps identify distinct sessions and the number of pages visited per session. Navigation pattern mining clusters sessions based on user behavior, revealing common navigation patterns and providing insights into user interactions with the website [45].

In our implementation, session identification relies on user identification. We employ the threshold time approach to determine the start and end of each session. By using an approximate value 30 minutes, we define a time threshold that separates repeated user interactions into distinct sessions. This allows to group user activities within a specific timeframe, making it easier to analyze and understand user behavior on the website. The selection of a session duration threshold, like 30 minutes, can vary based on the application or system's context and requirements. Factors like the website's nature, user behavior patterns, and desired session tracking granularity influence the choice. Although there is no universal standard, a 30-minute threshold is commonly used, assuming that user interactions within this timeframe constitute a single session.

However, the optimal threshold should be determined by evaluating specific use cases and system characteristics. It's important to fine-tune the threshold accordingly[46]. Each user's page duration is calculated by subtracting the time spent on the previous page from the time spent on the next page. In this log file, a session is considered complete after 30 minutes, and a new session starts immediately after[47].Table 6 provides an example of session data with user identification based on IP, OS, and browser, along with corresponding timestamps, referrer, and URL requests.

The table shows multiple sessions by the same user with IP address 109.93.18.182. Each session is assigned a unique session ID (S-ID) to differentiate between them. The timestamp indicates the time of each user's activity. The "Referrer" column shows the website or page from which the user arrived before accessing the current URL specified in the "URL Request" column. The last two columns, "OS" and "Browser," provide information about the user's operating system and browser, respectively.

Table 6: session (user identification by ip, os, browser) by time

IP Address	S-ID	Time	Referrer	URL Request	OS	Browser
109.93.18.182	1	11/16/2009 11:46:35 AM	-	/	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	1	11/16/2009 11:46:42 AM	http://www.vtsns.edu.rs/	/o_skoli.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	1	11/16/2009 11:46:44 AM	http://www.vtsns.edu.rs/o_skoli.php	/kontakt.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	1	11/16/2009 11:46:48 AM	http://www.vtsns.edu.rs/kontakt.php	/stud_sluzba.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	2	11/16/2009 6:47:30 PM	-	/	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	2	11/16/2009 6:47:33 PM	http://www.vtsns.edu.rs/	/oglasna.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	2	11/16/2009 6:47:43 PM	http://www.vtsns.edu.rs/	/oglasna.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	3	11/16/2009 9:59:40 PM	-	/	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	3	11/16/2009 9:59:52 PM	http://www.vtsns.edu.rs/	/vest.php?id=148	Windows NT 5.1	Firefox/3.5.5

The proposed new approach for session identification introduces a different criterion based on the referrer request. It states that a session should be considered as starting when the referrer URL is "-" (indicating that the user is arriving from a different website) and the URL request is "/" (representing the main page of the website).The reasoning behind this approach is to identify sessions where a user browses to a website from another source (indicated by a referrer of "-") and their initial interaction is with the main page of the website (URL request is "/"). This combination suggests a new session, as the user is entering the website from an external source and starting their engagement from the homepage. By considering this referrer and URL pattern, the approach capture meaningful session boundaries and differentiate between new visits to the website and subsequent interactions within an ongoing session.

Furthermore , It's important to note that the proposed approach assumes that a new session begins when the user accesses the website from an external referrer and starts their interaction from the main page. This assumption may need to be validated and fine-tuned based on the specific characteristics of the website, user behavior patterns, and the goals of the session identification process. More experimentation and analysis can help determine the effectiveness of this approach in accurately identifying sessions. Table 7 shows the three sessions associated with the IP address "109.93.18.182", the OS type "Windows NT 5.1", and the browser type "Firefox/3.5.5". These sessions are identified

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

based on the referrer information. the session data presented in Table 6 provides valuable insights into user behavior and website usage. Each row in the table represents a specific session for the IP address "109.93.18.182", operating system "Windows NT 5.1", and browser "Firefox/3.5.5". The sessions are identified based on the referrer and URL information.

Table 7: session (user identification by ip, os, browser) by referrer

IP Address	S-ID	Time	Referrer	URL Request	OS	Browser
109.93.18.182	1	11/16/2009 11:46:35 AM	-	/	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	1	11/16/2009 11:46:42 AM	http://www.vtsns.edu.rs/	/o_skoli.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	1	11/16/2009 11:46:44 AM	http://www.vtsns.edu.rs/o_skoli.php	/kontakt.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	1	11/16/2009 11:46:48 AM	http://www.vtsns.edu.rs/kontakt.php	/stud_sluzba.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	2	11/16/2009 6:47:30 PM	-	/	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	2	11/16/2009 6:47:33 PM	http://www.vtsns.edu.rs/	/oglasna.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	2	11/16/2009 6:47:43 PM	http://www.vtsns.edu.rs/	/oglasna.php	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	3	11/16/2009 9:59:40 PM	-	/	Windows NT 5.1	Firefox/3.5.5
109.93.18.182	3	11/16/2009 9:59:52 PM	http://www.vtsns.edu.rs/	/vest.php?id=148	Windows NT 5.1	Firefox/3.5.5

By examining the sessions' details, we can gain a deeper understanding of how users interact with the website. The timestamp of each activity allows us to analyze the duration of each session and identify patterns of user engagement. This information helps us determine the average length of user visits and evaluate the level of user interest and involvement.

The referrer column provides insights into the sources that drive traffic to the website. By examining the referrer URLs, we can identify external websites, search engines, or marketing campaigns that bring users to the site. This analysis helps in evaluating the effectiveness of different traffic sources and understanding user acquisition channels.

Additionally, studying the sequence of URL requests within each session reveals the paths users follow and the pages they visit. We can analyze the entry points where users first access the website and the exit points where they leave. This information is valuable for optimizing website design, improving user experience, and identifying popular or problematic areas of the site.

6.4 Path Complete

After determining a unique user session, it becomes important to identify significant page accesses that are not recorded in the log file due to caching mechanisms on the client or proxy side. When a user utilizes the back button in their browser to access a page, a cached copy of that page is retrieved instead of generating a new log entry. This lack of logging creates the issue of missing references, which requires the use of path completion techniques to fill in these missing entries in the log file[48]. Path completion involves determining the reference or source that was used to access a particular web page. It helps to determine whether web pages were accessed directly or through other reference pages. Not all requests can directly access the corresponding web page, so path completion is necessary to obtain the complete user access path. The incomplete access path of each user session is identified based on the identification of the user session[43].

In our implementation, Fig. 10 displays the outcomes of the path completion process for individual user sessions, categorized by time for IP Address, as well as by time for IP Address, OS, and Browser. The figure shows the results obtained from analyzing the complete paths of user sessions

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

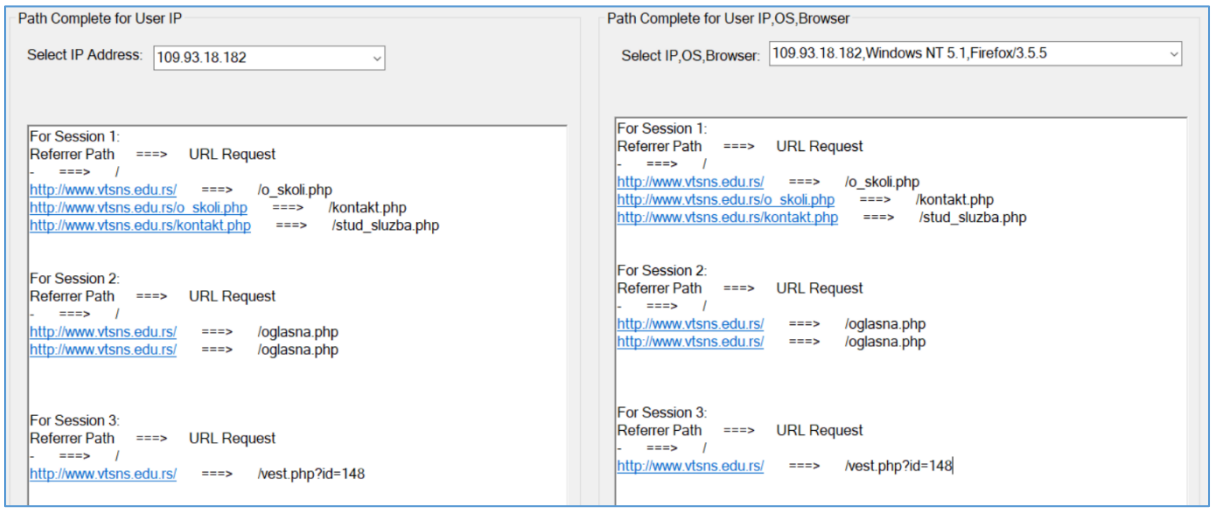

Fig. 10: Result of Path Complete step (Session Access Path Analysis by Time)

In contrast (Fig. 11)shows the analysis of path completion results (Our Approach) for user sessions grouped by referrer, with variations of IP Address, OS, Browser. It presents the results for each user session, with a specific focus on grouping them by referrer.

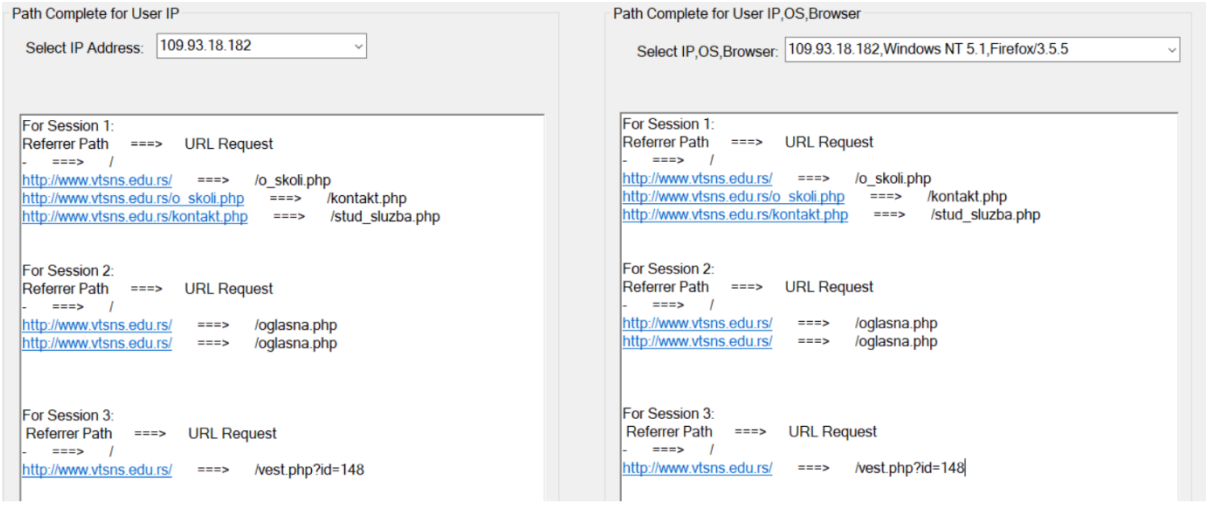

Fig. 11: Result of Path Complete step for session by referrer.

6.5 Statistical Analysis

in recent years, the website visitors is become more important, so Statistical analysis appears as a commonly used method. By examining the session data, various descriptive statistical analyses can be performed on variables like page views, viewing time, and navigational paths. For instance, e-Trade abandoned their German website as they discovered that German visitors were accessing the English site instead[49]. Web traffic analysis tools generate periodic reports that contain statistical information such as frequently visited pages, average page view duration, and average path length on the site. This knowledge can be valuable for improving system performance, enhancing security, facilitating website modifications, and aiding marketing decisions. Numerous commercial tools are available for showing statistical analysis in web analytics[50].

In our implementation, several statistical computations were conducted as part of our analysis. These computations are presented in (Table 8). Firstly, we determine the number of records in the dataset before and after the pre-processing stage. This helps to understand the impact of data cleaning and preparation on the dataset. Also, an approach were proposed to determine the device type used by visitors accessing the website. By examining the operating system (OS) associated with each request, we can categorize the device as either a computer or a mobile device. For computers, we

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

consider OS types such as Windows, Linux, and Mac, while for mobile devices, we include OS types like Symbian and 2ME/MIDP.

The dataset was analyzed to identify the number and names of different operating system types present. This allows to understand the diversity of operating systems used by website visitors. In cases where browsers like Safari are utilized, we leverage the OS information to differentiate between "Mobile Safari" and "Computer Safari" instances. The number of web pages was calculated based on their file extensions. This metric provides insights into the distribution of different types of web pages within the website. The number of requests made using the GET and POST methods was determined, while filtering out any other request methods during the data cleaning step. This will help to understand the occurrence of these two common request methods within the dataset.

Finally, the number of unique users or visitors who access the website was computed. This metric provides an understanding of the overall user base and can be valuable for various analytical purposes.

Table 8: statistical analysis result

Key	Values
ımber of Records in the Dataset.	Number of Record Before Preprocessing step = 5982 Number of Record After Preprocessing step = 1606
Device Types	Number of Computer = 1587 Number of Mobile = 19
Operating System Types	Number of Windows OS= 1561 Number of Linux OS= 24 Number of Mac OS= 2 Number of Symbian OS= 13 Number of J2ME/MIDP OS= 6
Browser Types	Number of Computer Firefox= 907 Number of Computer Flock= 7 Number of Computer Internet Explorer= 436 Number of Computer Opera= 156 Number of Mobile Opera= 6 Number of Mobile Safari= 13 Number of Computer Safari= 81
Page Extensions	Number of Home= 474 Number of php= 1127 Number of HTML= 4 Number of No Extension= 1
Request Methods	Number of GET= 1565 Number of POST= 41
Number of user (visitor)	Number of Users= 297
Number of pages	Number of Pages= 41 Page Name => No of requests (Visitors) / => 474 /konsultacije.php => 32 /vesti.php => 24 /oglasna.php => 207 /ispit_rezultati.php => 90 /raspored_predavanja.php => 33 /ispit_raspored_god.php => 43 /smerovi.php => 33 /nastavno_osoblje.php => 24 /profesor.php => 49 /ispiti.php => 124 /ispit_raspored_akt.php => 91 /nenastavno_osoblje.php => 6 /stud_sluzba.php => 10 /biblioteka.php => 5 /skriptarnica.php => 8

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

	/konkursi.php => 18 /smer_kurikulum.php => 41 /smer_raspored.php => 6 /specijalisticke.php => 23 /o_skoli.php => 20 /index.php => 22 /galerija.php => 58 /linkovi.php => 15 /kontakt.php => 15 /mapa.php => 11 /vest.php => 32 /studije.php => 3 /upis_prva.php => 14 /ispit_pravila.php => 4 /noviSajt/vtsns_admin/amfphp/gateway.php => 38 /noviSajt/vtsns_admin/amfphp/services/Godisnji.php => 1 /ispit_odbijeni.php => 11 /ispit_odbijeni_spisak.php => 9 /noviSajt/administrator/history/historyFrame.html => 4 /admin/ => 1 /admin/user.php => 2 /admin/index.php => 1 /admin/vesti.php => 1 /upis_ostale.php => 2 /cenovnik.php => 1

By performing these statistical computations, we gain valuable insights into various aspects of the website, such as device usage, operating system diversity, web page distribution, request methods, and user engagement.

7. Conclusion and Future Work

In the implementation step, the challenge is how to unstructured data with high noise in the access log file. There are two contributions were successfully made. firstly, determining the device type used (computer or mobile), this advantage are: optimizing advertising and recommendation systems to suit different device characteristics, personalized advertising and recommendation systems, and enabling businesses to tailor their strategies. And the second contribution is to identify user sessions and complete paths by leveraging the referrer URL, this advantage is: enables a deeper understanding of user behavior and facilitates pattern discovery. also, this paper proposes secure connection by determining robot URLs (such as search engines).

References

[1] P. Sukumar, L. Robert, and S. Yuvaraj, "Review on modern Data Preprocessing techniques in Web usage mining (WUM)," in *2016 international conference on computation system and information technology for sustainable solutions (csitss)*, 2016, pp. 64–69.

[2] P. Ristoski and H. Paulheim, "Semantic Web in data mining and knowledge discovery: A comprehensive survey," *J. Web Semant.*, vol. 36, pp. 1–22, 2016.

[3] K. Sharma, G. Shrivastava, and V. Kumar, "Web mining: Today and tomorrow," in *2011 3rd International Conference on Electronics Computer Technology*, 2011, vol. 1, pp. 399–403.

[4] [4] P. S. Sharma, D. Yadav, and R. N. Thakur, "Web page ranking using web mining techniques: a comprehensive survey," *Mob. Inf. Syst.*, vol. 2022, pp. 1–19, 2022.

[5] S. Yadao and A. Vinaya Babu, "Usage of Web Mining for Sales and Corporate Marketing," in *Communication Software and Networks: Proceedings of INDIA 2019*, 2021, pp. 55–60.

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[6] D. T. S. Bommi Harika, "Identification of User Behaviour by Web Usage Mining," *Math. Stat. Eng. Appl.*, vol. 71, no. 4, pp. 678–692, 2022.

[7] A. Kumar and R. K. Singh, "Web mining overview, techniques, tools and applications: A survey," *Int. Res. J. Eng. Technol.*, vol. 3, no. 12, pp. 1543–1547, 2016.

[8] J. Chaki, N. Dey, B. K. Panigrahi, F. Shi, S. J. Fong, and R. S. Sherratt, "Pattern mining approaches used in social media data," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 28, no. Supp02, pp. 123–152, 2020.

[9] J. Kapusta, M. Munk, and D. Halvoník, "Quality of Extracted Sequential Rules by Session Identification Using STT and Cookies," in *2017 European Conference on Electrical Engineering and Computer Science (EECS)*, 2017, pp. 150–154.

[10] S. Aggarwal and V. Mangat, "Application Areas of Web Usage Mining," in *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, 2015, pp. 208–211.

[11] Y. Bei and Z. Cai, "Online Extracting Sessions of Frequent Users," in *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, 2020, pp. 440–446.

[12] M. Udantha, S. Ranathunga, and G. Dias, "Modelling website user behaviors by combining the EM and DBSCAN algorithms," in *2016 Moratuwa Engineering Research Conference (MERCon)*, 2016, pp. 168–173.

[13] H. M. S. Lotfy, S. M. S. Khamis, and M. M. Aboghazalah, "Multi-agents and learning: Implications for Webusage mining," *J. Adv. Res.*, vol. 7, no. 2, pp. 285–295, 2016.

[14] Z. Ansari, S. A. Sattar, A. V. Babu, and M. F. Azeem, "Mountain density-based fuzzy approach for discovering web usage clusters from web log data," *Fuzzy Sets Syst.*, vol. 279, pp. 40–63, 2015.

[15] S. S. S. Ghaemmaghami, S. S. Emam, and J. Miller, "Automatically inferring user behavior models in large-scale web applications," *Inf. Softw. Technol.*, vol. 141, p. 106704, 2022.

[16] R. Mishra, P. Kumar, and B. Bhasker, "A web recommendation system considering sequential information," *Decis. Support Syst.*, vol. 75, pp. 1–10, 2015.

[17] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Appl. Comput. Informatics*, vol. 12, no. 1, pp. 90–108, 2016.

[18] J. P. Dias and H. S. Ferreira, "Automating the Extraction of Static Content and Dynamic Behaviour from e-Commerce Websites," *Procedia Comput. Sci.*, vol. 109, pp. 297–304, 2017.

[19] G. Shivaprasad, N. V. S. Reddy, U. D. Acharya, and P. K. Aithal, "Neuro-Fuzzy Based Hybrid Model for Web Usage Mining," *Procedia Comput. Sci.*, vol. 54, pp. 327–334, 2015.

[20] A. Ganibardi and C. A. Ali, "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*, 2018, pp. 263–271.

[21] S. Hooshmand, M. Faheem, G. V Bochmann, G.-V. Jourdan, R. Couturier, and I.-V. Onut, "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*, 2016, pp. 64–74.

[22] M. Srivastava, R. Garg, and P. K. Mishra, "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), 2015.

[23] S. Samiei, M. Joodaki, and N. Ghadiri, "A scalable pattern mining method using apache spark platform," in *2021 7th international conference on web research (ICWR)*, 2021, pp. 114–118.

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[24] S. Sulaimany and A. Mafakheri, "Visibility graph analysis of web server log files," *Phys. A Stat. Mech. its Appl.*, p. 128448, 2023.

[25] R. Roy and G. A. Rao, "Survey on pre-processing web log files in web usage mining," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 3 Special Issue, pp. 682–691, 2020.

[26] K. Mani and K. R. Suneetha, "Performance evaluation of Compact Prediction Tree algorithm for Web Page Prediction," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1–7.

[27] J. Svacina *et al.*, "On vulnerability and security log analysis: A systematic literature review on recent trends," in *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, 2020, pp. 175–180.

[28] C. Christner, A. Urman, S. Adam, and M. Maier, "Automated tracking approaches for studying online media use: A critical review and recommendations," *Commun. Methods Meas.*, vol. 16, no. 2, pp. 79–95, 2022.

[29] A. S. Rajawat and A. R. Upadhyay, "Web personalization model using modified S3VM algorithm for developing recommendation process," in *2nd International Conference on Data, Engineering and Applications (IDEA)*, 2020, pp. 1–6.

[30] S. Sharma and Dalip, "Performance Evaluation of Secure Web Usage Mining Technique to Predict Consumer Behaviour (SWUM-PCB)," in *Intelligent Computing and Networking: Proceedings of IC-ICN 2021*, Springer, 2022, pp. 136–145.

[31] G. Weng *et al.*, "HawkDock: a web server to predict and analyze the protein--protein complex based on computational docking and MM/GBSA," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W322--W330, 2019.

[32] P. Buono, F. Balducci, and others, "MonitorApp: a web tool to analyze and visualize pollution data detected by an electronic nose," *Multimed. Tools Appl.*, vol. 78, no. 23, pp. 33023–33040, 2019.

[33] M. Srivastava, A. K. Srivastava, and R. Garg, "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*, 2019.

[34] "Higher Education Technical School of Professional Studies in Novi Sad," 2022. [Online]. Available: https://vtsns.edu.rs/about-the-school/. [Accessed: 03-Jul-2023].

[35] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," *Int. J. Inf. Technol.*, vol. 15, no. 2, pp. 965–980, 2023.

[36] A. Asif, D. AlFrraj, and M. A. Alshamari, "A comprehensive approach of exploring usability problems in enterprise resource planning systems," *Appl. Sci.*, vol. 12, no. 5, p. 2293, 2022.

[37] L. Berti-Equille, "Learn2clean: Optimizing the sequence of tasks for web data preparation," in *The World Wide Web Conference*, 2019, pp. 2580–2586.

[38] J. Zhou, J. Wei, and B. Xu, "Customer segmentation by web content mining," *J. Retail. Consum. Serv.*, vol. 61, p. 102588, 2021.

[39] R. Manikandan and V. Saravanan, "A novel approach on Particle Agent Swarm Optimization (PASO) in semantic mining for web page recommender system of multimedia data: a health care perspective," *Multimed. Tools Appl.*, vol. 79, pp. 3807–3829, 2020.

[40] A. S. Albahri *et al.*, "Multi-biological Laboratory Examination Framework for the Prioritization of Patients with COVID-19 Based on Integrated AHP and Group VIKOR Methods," *Int. J. Inf. Technol. Decis. Mak.*, pp. 1–23, 2020.

[41] R. Patil and P. Trivedi, "Determination of User Navigational Patterns from Server Log Files using Hadoop Techniques."

Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 2, 2024, pp. 54-74
DOI: https://doi.org/10.25195/ijci.v50i2.486
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[42] "About /robots.txt," 2007. [Online]. Available: http://www.robotstxt.org/robotstxt.html.

[43] V. Leno, A. Polyvyanyy, M. Dumas, M. La Rosa, and F. M. Maggi, "Robotic process mining: vision and challenges," *Bus. \& Inf. Syst. Eng.*, vol. 63, pp. 301–314, 2021.

[44] F. Masood *et al.*, "Spammer detection and fake user identification on social networks," *IEEE Access*, vol. 7, pp. 68140–68152, 2019.

[45] M. de Leoni and S. Dündar, "Event-log abstraction using batch session identification and clustering," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 36–44.

[46] "Average session duration in Google Analytics." [Online]. Available: https://www.hotjar.com/google-analytics/glossary/session-duration/.

[47] S. Asadianfam, H. Kolivand, and S. Asadianfam, "A new approach for web usage mining using case based reasoning," *SN Appl. Sci.*, vol. 2, no. 7, p. 1251, 2020.

[48] W. Wei, Q. Ke, J. Nowak, M. Korytkowski, and M. Scherer Rafałand Woźniak, "Accurate and fast URL phishing detector: a convolutional neural network approach," *Comput. Networks*, vol. 178, p. 107275, 2020.

[49] A. Mohamed, M. K. Najafabadi, Y. B. Wah, E. A. K. Zaman, and R. Maskat, "The state of the art and taxonomy of big data analytics: view from new big data framework," *Artif. Intell. Rev.*, vol. 53, pp. 989–1037, 2020.

[50] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data Cogn. Comput.*, vol. 4, no. 1, p. 1, 2020.