International Journal of e-Collaboration
Volume 19 • Issue 1

Web Mining Techniques - A Framework to Enhance Customer Retention

Shimaa Ouf, Helwan University, Egypt*
Yehia Helmy, Helwan University, Egypt
Merna Ashraf, Helwan University, Egypt

ABSTRACT

In e-commerce, retaining customers on the web is a difficult task that requires a good understanding of customers' behavior to be able to predict their needs and interests. Web usage mining (WUM), which is the application of data mining techniques to improve business, helps in understanding customers' behavior on the web. Therefore, this paper proposes and implements a framework to enhance the quality of customer recommendations. Providing customers with what they are looking for helps increase their satisfaction, which will lead to improved retention with the company. The proposed framework was tested and evaluated. The result of testing the proposed framework illustrates that the recommendations based on merged techniques (like clustering, classification, association, and sequential discovery) achieve strong accuracy with a precision value of 74%, coverage of 100%, and an average overall efficiency of F-measure of 86%. which means that the merged technique outperformed each technique and attained much higher overall coverage.

KEYWORDS

Competitive Advantage, Customer Relationship Management, Data Mining, Recommendation Layer, Software Architecture, Web Server Log, Web Usage Mining

INTRODUCTION

E-commerce is growing exponentially in terms of business and data. Many organizations became highly reliant on their websites to attract new customers and retain the existing ones (Lopes & Roy, 2015). The rapid expansion has resulted in many challenges for organizations such as retaining their customers in this highly competitive environment. Multiple choices for each product or service were available to customers, so their decision to shift from one organization to another became very easy. Also, Organizations have difficulty in how to benefit from this enormous data in retaining their customers and increasing their revenues. Therefore, organizations need to adopt a marketing strategy such as customer relationship management (CRM). CRM is an approach that stems from the need to create a new business environment, which allows for more effective management of the relationship with customers. CRM is the process of acquiring, retaining, and collaborating with selected customers to create superior value for both company and the customer.

DOI: 10.4018/IJeC.315790 *Corresponding Author

In this highly competitive market, retaining customers gives greater benefits to organizations than acquiring new ones. Research has shown that companies that create satisfied, loyal customers have more repeat business, lower customer-acquisition costs, and stronger brand value. All of which translates into better financial performance (Soltani & Navimipour, 2016). The effective way to increase customer retention in e-commerce is to provide the customers with recommendations that match their preferences and interest, this requires a good understanding of customers' (user) behaviors on the web, which led to the emergence of the web usage mining concept (V. Kumar & Ayodeji, 2021).

Web Usage Mining (WUM) is the application of data mining techniques to discover interesting usage patterns from web data (Chavda, Jain, Panchal, & Valera, 2017a). It is the process of extracting useful information from various weblogs. WUM focuses on predicting customers' (users') preferences and behaviors by analyzing their behavior on the web (A. Kumar, Bhushan, Pokhriya, Chaganti, & Nand, 2022).

The role of web mining in the CRM is to extract useful patterns from users' behaviors that can be used in providing recommendations to them related to their preferences and interests which will result in increased customers' loyalty and therefore their retention which is the aim of the paper.

WEB USAGE MINING

A large amount of data available on the site's web pages made the organizations and companies focus on gathering this data. This data is used for many purposes such as predicting users' interests and needs. Web prediction is a field of web usage mining in which users' future behavior on the web is predicted (Mittal, Malik, Rattan, & Jhamb, 2021). The results of the prediction can be used for:

- Personalization of web content.
- Reducing the server response time (Sellamy et al., 2018).
- Provide guidelines for improving the design of web applications.

Data Sources of Web Usage Mining

Web usage mining is used to extract users' access patterns to track the user's browsing behavior which helps to understand users' interests and preferences (J. Kaur & Garg, 2019). Web usage mining highly depends on data collected from Web Server Side, Proxy-Side, and Client-Side. Which are known as usage data, the usage data is data collected automatically by the web which represents the fine-grained navigational behavior of users (Neelima & Rodda, 2016). Table 1 illustrates the three sources of data.

Table 1. Types of Usage data

Criteria	Web Server-side	Client-side	Proxy side
Definition	Data that are collected from web servers.	Data are collected from the host that accesses the website.	Data are collected from the Proxy server which is a software system that acts as an intermediary between the client and other servers (Aldekhail, 2016).
Participants	Multi-users	Single user	Multi-users
Websites	Single site	Multi-site	Multi-site
Reliability of data	Due to caching and network transmission times, data may not be entirely reliable.	More reliable than server data because they avoid caching and IP misinterpretation problems.	Low reliability due to caching and IP misinterpretation.
Source of data	Data accessed from: • Server log file • Cookies • Explicit user input	Data are accessed through sending remote agents implemented in Java or JavaScript and then embedded in web pages.	Data is accessed by using access logs to store web page requests and responses from servers.

The purpose of this paper is to understand the behavior of users of an e-commerce website to help provide personalized recommendations for them to increase their satisfaction and then increase their retention of this site.

Because sites that use predominantly dynamic content such as e-commerce sites, will not be affected by proxy-level caching for each page served is technically a new file (Selvy, Anitha, Varthan, Sethupathi, & Adharsh, 2022), the proxy side was ignored. And the client-side because it focuses on the behavior of the single user on multiple sites which is not our purpose and considers only the case of server-side and especially how to use a web server log file.

Web Server Log File

The web server log is a fundamental data source in WUM. It explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the concurrent and interleaved access to a website by multiple users.

There are many types of web server log files. Each of them contains specific data that performs a specific task. It is very important to determine which type of data you are looking for and which data fields help you to accomplish your goal:

- **Access log:** It stores information about which files are requested from the web server, and all the clicks, hits, and accesses made by the web users (Nguyen, Diep, Hoang Vinh, Nakajima, & Thoai, 2018).
- **Referrer log:** It stores information about the URLs of the web pages on other sites that link to the site's web pages. If a user gets to one of the server pages by clicking on a link from another site, the URL of that site will appear in this log.
- **Agent log:** It records information about the web clients that send requests to web servers (Britvin, Alrawashdeh, & Tkachuk, 2022). Contain type of browser and the platform determines what a user can access the website.
- **Error log:** It stores information about errors and failed requests on the web server. Web log files can be stored and written in various formats. Each of these formats has its own fields.

The most used log formats by many researchers are Common Log Format (CLF) and Extended Common Log Format (ECLF). An example of CLF is shown in Table 2.

The most used log formats by many researchers are Common Log Format (CLF) and Extended Common Log Format (ECLF) (Ganibardi & Ali, 2018). An example of CLF is shown in Table 3.

In this paper, only CLF for accessing web log files is taken into consideration because it contains the required data that is used to identify users' preferences and interests, this data will then be used to provide users with personalized recommendations. The basic data fields that exist in the access log file in CLF are:

- The client IP address
- User ID
- Date/time
- Request (method, URL, protocol)
- Status code
- Bytes
- Referrer
- Agent

The log file entries produced in CLF will look like Figure 1. Table 4 gives a full description and example of data fields in CLF.

Table 2. Weblog file formats

Criteria	Common Log Format	Extended Common Log Format	Internet Information Services log Format
	CLF	**ECLF**	**IIS**
Description	Is a standardized text file format used by web servers when generating server log files.	Is a customizable ASCII format, with a variety of different fields that are used by web servers, when generating log files.	Is a fixed ASCII text-based format that cannot be customized. The IIS log file contains the HTTP Server API kernel-mode cache hits (Neelima & Rodda, 2016).
Advantage	The format is standardized so that analytic programs can more conveniently make use of the information contained within them.	Provide more information and flexibility than CLF files.	Contains the most information.
Meaning of record	Each line represents a single user request.	Each line can contain either a direct or an entry. Entries consist of a sequence of fields relating to a single HTTP transaction. Directives record information about the logging process itself.	Each line represents a single user request.
A log' record might contain	IP address, User ID, date/time, request (method, URL, protocol), status code, bytes, referrer, and agent.	Version, Fields, Software, Start date, End date, Date, Remark.	Client IP address, Username, Date, Time, Service and instance, Server name, Server IP address, Time taken, Client bytes sent, Server bytes sent Service status code, Windows status code, Request type, Target of operation, Parameters.

Table 3. Sample of cleaned Common Log Format (CLF)

IP address	User ID	Date/Time	Method	URL	Protocol	Status code	Bytes
10.1.109.54	-	[15/ Jul/2010:21:24:14]	GET	/displaytitle. php?id=10	HTTP/1.1	200	3024
10.1.11.81	-	[15/ Jul/2010:21:24:17]	GET	/displaytitle. php?id=10	HTTP/1.1	200	4029
10.1.11.81	-	[16/ Jul/2010:02:51:41]	GET	/mail/mailgraph.png	HTTP/1.1	200	2056
10.1.11.81	-	[16/ Jul/2010:03:00:21]	GET	/about-us/	HTTP/1.1	300	5486
10.1.11.81	-	[16/ Jul/2010:03:00:25]	GET	/release-schedule/	HTTP/1.1	404	3582
10.1.11.81	-	[16/ Jul/2010:03:00:29]	GET		HTTP/1.1	200	1256
10.1.114.150	-	[16/ Jul/2010:03:00:31]	GET	/faq/	HTTP/1.1	200	2478

Figure 1. A sample of web access log entries

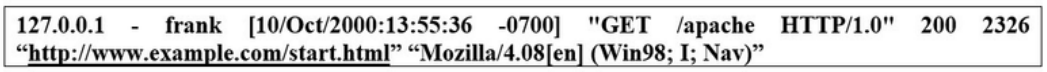


Table 4. CLF' data fields

Data field	Description	Example
IP address	The IP address of the client is requested to the server (Nguyen, Diep, Hoang Vinh, et al., 2018).	127.0.0.1
User authentication	Username and password if the server requires user authentication (generally empty and represented by a "-") (Shivaprasad, Reddy, & Acharya, 2015)	"-"
User ID	Is filled in only when authentication is required to access the files.	Frank
Date/Time	Used to determine how long a visitor has spent on a given page (Shivaprasad et al., 2015).	[10/Oct/2000:13:55:36 -0700]
Request	The request field contains three important information which are Method: GET; URL: apache_pb.gif; Protocol version: HTTP/1.0.	"GET /apache_ HTTP/1.0"
Status code	Is sent by the web server and indicates the action taken in response to the request.	200
Bytes	The content length of the document was transferred.	2326
Referrer	Contain the URL of the source of the request.	"http://www.example.com/start.html"
Agent	Provide information about the client's browser, the browser version, and the client's operating system.	"Mozilla/4.08[en] (Win98; I; Nav)"

The web access log entry in Figure 2 shows that a user from the IP address 127.0.0.1 successfully requested the page "apache" on 10/Oct/2000 at 13:55:36p.m. The user has used the "Mozilla" browser, "Win98" operating system, the HTTP protocol, and method "GET" to access the page and a total of 2326 bytes were returned. Generally, the log file is collected in its raw form, so it contains a lot of irrelevant, noisy, incomplete, duplicate, and missing data which restrict the identification of precise usage patterns with it (Patel & Parikh, 2017), as shown in Figure 3.

To clean the data, Data preprocessing is the most important phase of Web Usage Mining which is responsible for cleaning data and preparing it for the pattern discovery phase (J. Kaur & Garg, 2019).

Data Preprocessing Phase

After the raw weblog data were collected from the server-side as mentioned above, these data must be transformed into processed data, which is the role of the data-preprocessing phase (J. Kaur & Garg, 2019).

Data pre-processing phase is the most complex phase, it may take 80% of the mining process. Data pre-processing consists of many processes (Nguyen, Diep, Vinh, Nakajima, & Thoai, 2018).

Each of these processes has an essential role to prepare the raw data into structured data that can be used in the next phase of WUM which is the pattern discovery phase. The successful application of data mining techniques in pattern discovery phases are highly dependent on the correct application of the preprocessing phase (Saleh Ibrahim et al., 2022). Table 5 illustrates the role of each process.

Figure 2. Raw log file

Figure 3. Data preprocessing processes (Adeniyi, Wei, & Yongquan, 2016)

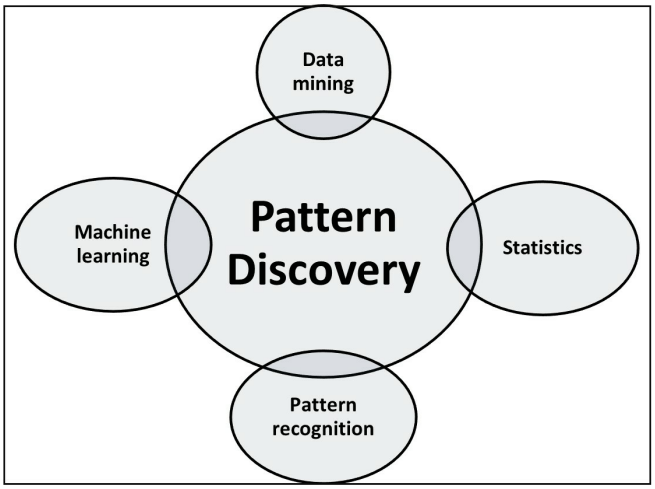
Table 5. The role of data preprocessing processes

Criteria	Data Preprocessing processes		
	Data cleaning	**User identification**	**Session identification**
Main task	Is responsible for removing irrelevant, duplicated, missing value and noise data (Patel & Parikh, 2017).	Is responsible for identifying each distinct user.	Is responsible for identifying the sequence of activities performed by the user from the moment he enters the website to the moment he leaves the website (Shivaprasad et al., 2015).
Methods of process	By removing graphics files. By removing failed requests. By removing robots' requests.	By IP address. By user authentication. By cookies. By client information. By site topology.	By the time gap. By referral attribute. By time spent on observing page.

Pattern Discovery Phase

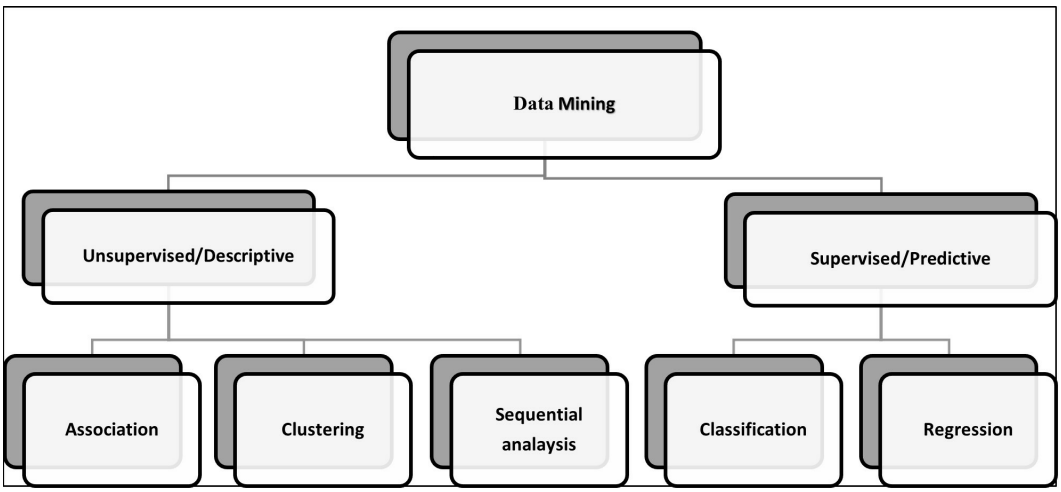
Once the preprocessing tasks have been completed successfully, the pattern discovery phase can be applied to the usage data. Pattern discovery is the second phase in WUM (Sathya & Devi, 2017) which draws upon methods and algorithms developed from several fields, as shown in Figure 4.

Figure 4. The components of the pattern discovery phase (Tiwari, S., et al. 2019)



The pattern discovery phase applies data mining techniques to discover useful patterns (Samboteng, Rulinawaty, Kasmad, Basit, & Rahim, 2022). The purpose of this phase is to extract knowledge (Kaadoud, Rougier, & Alexandre, 2022) such as extracting users' navigational patterns, identifying who is visiting the site, the path visitors take through the pages, and how much time visitors spend on each page, The most common starting page, and where visitors are leaving the site. These patterns are extracted from the preprocessed data by using data mining techniques like association, clustering, classification, and sequential discovery (Sathya & Devi, 2017). DM techniques are broadly divided into two categories: unsupervised/Descriptive and supervised/Predictive learning techniques (Hariharakrishnan, Mohanavalli, & Kumar, 2017), as shown in Figure 5.

Figure 5. Data mining techniques (Neelima, G. and S. Rodda., 2015)



Association Rule

In web usage mining, association rules refer to a set of pages that have been accessed together with a minimum support value, this can help in providing the right recommendations to the users (Dogan, Kem, & Oztaysi, 2022).

Clustering

Clustering is used to divide data into clusters that differ from each other but are similar from the inside (Zhu). There are two kinds of interesting clusters that can be discovered in web usage mining, User clustering which finds users having similar browsing behavior, and Page clustering which finds pages of similar content (Cai, Wang, Jiang, Zhang, & Peng, 2022). Clustering users' records (sessions or transactions) means establishing groups of users exhibiting similar browsing patterns (Zhang, Lin, Lin, & Liu, 2016).

Sequential Discovery

Sequential discovery in web usage mining is used to find users' navigation patterns and predict users' next visit (Aldekhail, 2016). For personalization and to provide high-quality recommendations to users sequential modeling is based on stochastic methods on the sequences of pages in user sessions to learn probabilistic models that can be used for predicting subsequent visits. One such approach that models the navigational activity on the Web site is the Markov model (Roy, 2021).

Classification

Classification in a log file gives useful rules such as 50% of users who purchase online lived downtown. To generate recommendations based on the classification technique (Panigrahi et al., 2021), the user is classified into a group that has users with similar behavior. The K-nearest neighbors (KNN) algorithm is considered the most useful technique in this domain (Zidan et al., 2022).

Pattern Analysis Phase (Recommendation Phase)

Pattern Analysis is the final phase of WUM, the discovered patterns are further processed and filtered to be used as input to applications such as dynamic recommendation, web analytics, and report generation tools (Sirichanya & Kraisak, 2021). Here pattern analysis phase is the recommendation phase. The recommendation phase is responsible to recommend or predicting what kind of items the user may prefer. Based on the data collected from the system's observed activities of the user (Moreno, Segrera, López, Muñoz, & Sánchez, 2016).

SYSTEMATIC LITERATURE REVIEW

Monitoring and understanding how web usage mining is used in recommender systems to increase customer satisfaction which will lead to enhanced customer retention is an active area of research in both the academic and commercial worlds.

Research Methodology

The purpose of this paper is to understand, how, web usage mining and its phases are used in understanding users' behavior to provide high-quality recommendations to the user based on his/her behavior that will result in increased customer retention which is the core of CRM. The following electronic journal databases, Springer Link, IEEE Library, and Science Direct are used to find relevant literature that fulfills the purpose of the paper. Only peer-reviewed articles written in English will be selected. Conference papers, masters' and doctoral dissertations, textbooks, unpublished working papers, and non-English papers were eliminated. The ambition is to use literature published between "2015 to 2021". The paper focuses on how to increase customer retention by understanding their

behavior and providing them with accurate recommendations. The search was performed based on the following search terms "CRM and Web usage mining", "CRM and personalized recommendation", and "Data mining and recommender system". The results of the search process were 221 articles.

Classification of Research Papers

The selected papers are categorized based on the scientific databases and the publication year.

- **Classification by scientific databases:** The research papers are chosen from many scientific databases, but most of them are chosen from the three most popular databases. Springer Link introduced (103 out of 221 research papers), Science Direct (92 out of 221 research papers), and IEEE library (26 out of 221 research papers), classification of research papers for the most popular scientific databases is shown in Figure 6.

LITERATURE REVIEW

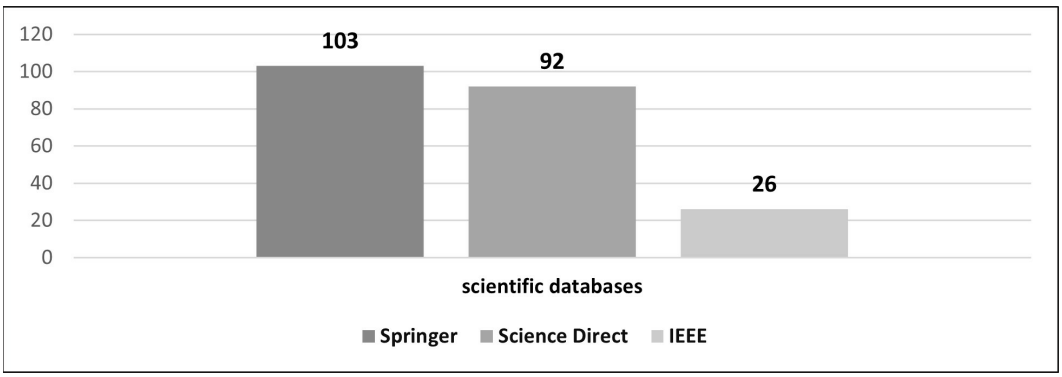
With the growing popularity of the World Wide Web, millions of users access websites all over the world. When the users browse the web pages, a large amount of data is recorded in the weblog files. Many academic researchers have been done on web usage mining in the prediction of user browsing behavior to improve customer retention.

(N. Kaur & Aggarwal, 2015) mentioned that web usage mining is very important for customer relationship management because it ensures customer satisfaction and the interaction between customers and companies. So, they applied the Weblog Expert Lite 8.6 tools to the weblog file to analyze it. The results helped to identify the most visited pages, most downloaded files, most used browsers, and most used operating systems which were used to predict user behavior on the website and customize the website to be user-friendly.

(Aldekhail, 2016) mentioned that the usage data resulting from the interaction between the users and the internet is very useful due to the huge amount of information that it provides about users' preferences, using web usage mining helps in analyzing these usage data which leads to extracting many decisions related to marketing. He explained and applied Web usage mining to find useful patterns that identify users' preferences.

(Ganibardi & Ali, 2018) introduced a rule-based cleaning method to clean the web log files from the noise. They mentioned that their method focused on the structure of the website in comparison to content-centric filtering heuristics are based on the requested resources. The results showed that the

Figure 6. The classification of the research paper by scientific databases



rule-based method has a significant advantage over the content-centric method for weblog cleaning concerning relevancy, workability, and cost constraints.

(Thorat, Goudar, & Barve, 2015) mentioned that due to the importance of effective communication between businesses and users, companies need to have a website that satisfies the needs of their users. They analyzed the webserver log that contains information about the usability of users to the website. They found that the weblog contains many noisy and redundant data. They proposed three algorithms that help in cleaning the data and identifying users and their sessions. The result showed that the algorithms help to clean data effectively, identify a user and session and provide information to study the interesting patterns of users.

(Shivaprasad et al., 2015) discussed the role of a recommender system in business to consumers of the e-commerce website. They identified the three stages of the recommender system in the domain of e-commerce which are understanding consumers, delivering recommendations, and the impact of the recommender system in providing personalized offerings to individual consumers. They also provided a comprehensive view of the personalization process includeing both the system side and the consumer side.

(Sathya & Devi, 2017) described the different processes used to prepare the data for the pattern discovery phase. They applied data cleaning, user identification, and session identification, then implemented the A priori algorithm on these data to extract the users' navigational behavior which results in identifying a set of frequent links in the data.

(Suchacka & Chodak, 2017) mentioned the problem of understanding the behavior of customers of an e-commerce website by analyzing the web log file. They applied association rule mining to an online bookstore and the results allowed them to formulate prediction rules that identify whether the visited user will purchase or not.

(Srinivas, 2017) mentioned that the data stored in weblog files merged with many eroded, incomplete, and unnecessary information. The original log file cannot be used directly to discover useful patterns. He presented different techniques to reprocess data, such as data cleaning, data fusion, and data integration. To remove irrelevant status codes, robots, and graphics files.

(Sengottuvelan, Lokeshkumar, & Gopalakrishnan, 2017) mentioned that identifying customers' (user) purchasing patterns for the website owner is a complex task because there is no direct relationship between data collected in the weblog files and processes that customers perform to purchase something. As a result, the users' sessions were used to identify customers who purchase from the site. They applied classification algorithms to find information about the user's behavior to predict potential buyers of an e-commerce website.

They proposed a new technique that merges between association rule algorithm and logistic regression. To build a model that predicts whether the customer was with a product or not. The results show that this model works more efficiently and has higher accuracy than logic or AR mining used standalone.

(Bandyopadhyay, Thakur, & Mandal, 2017) proposed a priori algorithm-product recommendation framework for recommending products to users based on the relationships between the products that exist in the users' sessions. The results showed that 60% of the recommended rules are correct.

(Subramaniyaswamy & Logesh, 2017) proposed a prediction model that uses a new variant of the algorithm as adaptive KNN for a collaborative filtering recommender system that correlates the user preferences and features of items for user modeling. The results proved the better performance of the proposed AKNN algorithm over other algorithms when highly sparse data for recommendation generation were considered.

(Adeniyi et al., 2016) introduced a study on web usage mining and recommender system that used the implicit data gathered automatically by the server. They applied the KNN algorithm to the current user session to classify sessions, into the most similar group. The results showed that the KNN classifier is transparent, consistent, straightforward, simple, and very helpful in classifying active sessions.

(Singh, 2020) proposed a model-based recommender system that can overcome the problems of scalability and sparsity. The proposed model applied the clustering technique to reduce these problems. Different clustering algorithms are implemented in the data to find the advantage and shortcomings of each algorithm in the recommendation domain and the results showed that k-means has high accuracy and is widely used in recommender systems.

Table 6 summarized the data mining techniques used by researchers to enhance the recommendations offered to customers in the domain of e-commerce. Most researchers used two or more techniques, but no one used them all to improve recommendations offered to users and overcome is shortcomings of each technique. Therefore, this paper proposed a framework that integrates the four techniques to enhance the quality of recommendations offered to customers and provide them with the most similar things that match their preferences and interests. Providing customers with what they are looking for helps increase their satisfaction which will lead to increasing their retention to the company that provides them with what they need. This positively affects the company by increasing its revenues.

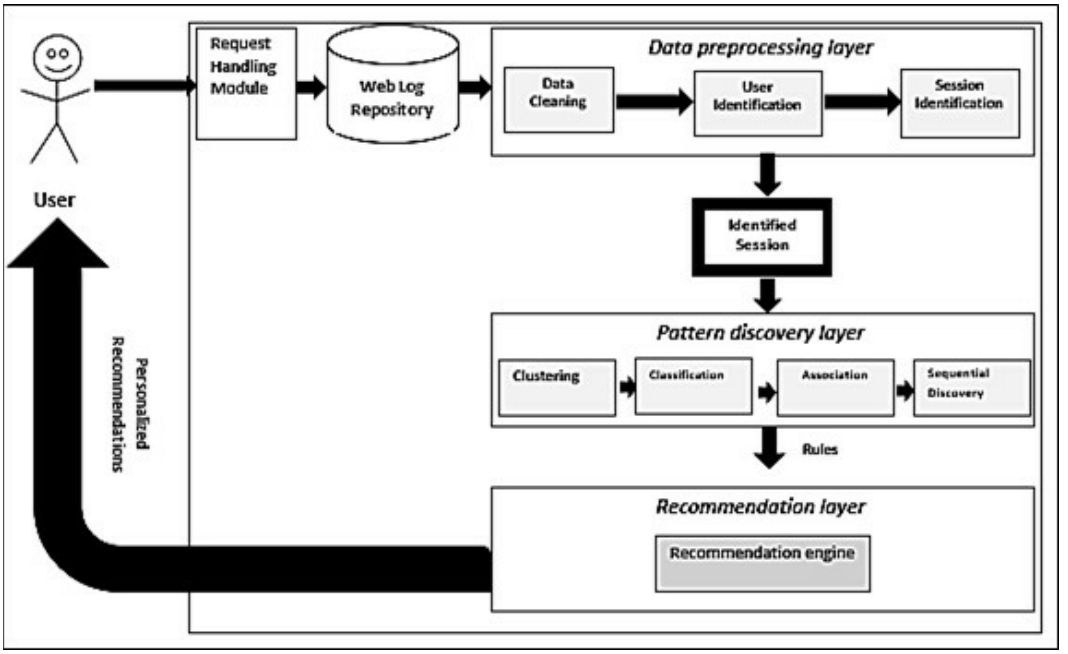
PROPOSED FRAMEWORK

The proposed framework consists of 3 layers. The first layer is the data preprocessing layer. The second layer is the pattern discovery layer, and the third layer is the recommendation layer. Figure 7 presents the layers of the framework that will provide more accurate recommendations to users. The data mining techniques such as clustering, classification, association rule, and sequential discovery in the pattern discovery layer, are used to track user behavior to provide him/her with the best recommendations. The following subsections discuss the framework's layers.

Table 6. A summary of the data mining techniques used by some researchers

Author	Sequential discovery	Association	Clustering	Classification
(Ismail, Ibrahim, Sanusi, Nat, & Sciences, 2015)	×	✓	×	✓
(Bahari & Elayidom, 2015)	×	×	×	✓
(N. Kaur & Aggarwal, 2015).	×	✓	✓	×
(Isinkaye, Folajimi, & Ojokoh, 2015)	×	✓	✓	✓
(Elhebir, Abraham, & Computing, 2015)	×	✓	✓	×
(Sengottuvelan, Lokeshkumar, & Gopalakrishnan, 2015)	✓	✓	✓	×
(Swamy, Babu, VENKATASUBBAIAH, & Research, 2015)	×	✓	×	×
(Soltani & Navimipour, 2016)	×	×	✓	✓
(Neelima & Rodda, 2016)	×	✓	✓	✓
(Chavda, Jain, Panchal, & Valera, 2017b)	×	✓	✓	×
(Sathya & Devi, 2017)	×	✓	×	×
(Hao, Zhaoxiang, & Bingbing, 2017)	×	×	✓	×
(Mehra & Thakur, 2018)	×	✓	✓	✓
(Nguyen, Diep, Vinh, et al., 2018)	×	✓	✓	✓
(Tiwari, Gupta, & Kashyap, 2019)	✓	×	✓	×
Proposed Framework	✓	✓	✓	✓

Figure 7. A framework for personalized recommendations



Request Handling Module

It handles the request made by the customer and logged the requests in a weblog file.

Weblog Repository

The web server log is a fundamental data source in WUM (Mehra & Thakur, 2018). It explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the concurrent and interleaved access to a website by multiple users. As mentioned before, there are many types of web server log files, each of which contains specific data that performs a specific task (Nguyen, Diep, Hoang Vinh, et al., 2018). Here, the focus is on the access log file that records the access information of users. Also, as mentioned before that Web log files can be stored and written in various formats (Ganibardi & Ali, 2018), where the focus will be on the log file in common log format which includes information about the user and the request made by him/her such as IP address, User ID, date/time, request (method, URL, protocol), status code, bytes, referrer, and agent. Every single line in the web log file represents a single user.

Data-Preprocessing Layer

Data pre-processing is the most important layer of web usage mining to clean web log files. The log file suffers from various data anomalies such as irrelevant, inconsistent, and noisy data. The purpose of this layer is to remove these anomalies and to transform data into a structured format which is used later to apply the different data mining techniques. Data pre-processing plays an important role in increasing the accuracy of recommendations which means that the higher the cleaning of data, the more the accuracy of the recommendations. The data-preprocessing layer consists of sub-processes which are:

1. Parsing web server access log file.
2. Significant attribute selection.

3. Data selection.
4. Data cleaning.
5. User identification.
6. Session identification.

Parsing Web Server Access Log File

Parsing the log's fields is the process that responds to separate the access log file' fields from a single entry into multiple fields. Data are stored in the access log file in an unstructured format, so by parsing data, the data are transformed into fields and rows.

Significant Attributes Selection

This process is responsible to select significant attributes such as IP address, date/ time, request field (method/URL/protocol), and status code. These attributes provide valuable information about the users' behavior while fields such as User ID are ignored because it contains "**-**" which means that it is empty and the Bytes field which is not adding value.

Data Selection

When a Web server receives an HTTP Request, it returns an HTTP response code to the client. This HTTP status code is a three-digit number (Kaur and Garg., 2019). All records with status codes below 200 and above 299 are not used for analysis, so this process is responsible to remove all records that have a status code below 200 or above 299 which means an error or failure.

Data Cleaning

It is the step in the data pre-processing layer that is responsible for removing irrelevant and noise records that contain Duplicated data, Missing values, unwanted graphics files, Non-GET methods, and Robots. So, by the end of this step, all this noise will be eliminated.

User Identification

This step is responsible for identifying each distinct user. In this paper, an e-commerce website is used, and all users are in a local network, so each user has a specific IP address without the agent's disturbance (Jardine, 2021). We identify users by assigning each user to a unique IP address in the log file.

Session Identification

The session identification step is responsible to identify the user's navigational behavior. Here, the user's session is identified by the Time gap heuristics in which a set of pages is considered as a single user session if those pages are requested within a specified period, which is 30 minutes and if the request was given after that threshold, it is considered as a new request (session) (Bayir & Toroslu, 2022). At the end of this layer, the data became ready to be used in the pattern discovery phase.

Pattern Discovery Layer

In this layer, the pre-processed log file is analyzed to extract valuable patterns by using data mining techniques. The most-known data mining techniques used are clustering, classification, association rule, and sequential discovery.

Clustering on the Weblog File

A user transaction clustering is applied, which is used to establish groups of users exhibiting similar browsing patterns. These clusters are used to create aggregated usage profiles, each profile is then represented as a vector in the original n-dimensional space. These aggregate profiles are used directly

to provide recommendations to new users. The main goal of applying user clustering is to analyze each segment of users and use these data to provide content to users that are personalized to their interest through the recommendations offered to them.

Classification on the Weblog File

Classification is the task of mapping data into one of several classes. A classification algorithm such as K-nearest neighbor (KNN) was used to measure the correlation between the current user profile (active session) and the past aggregated profiles resulting from the clustering technique to find an aggregate profile in the database like the preferences or characteristics of the target user.

Association Rule on the Weblog File

The association rule methods such as the Apriori algorithm is used to find groups of pages that frequently occur together in many transactions. These frequent itemsets are going to be used in the recommendation layer to provide effective recommendations to the target users. When a target user enters the website his/her preferences are matched against the antecedent **x** of each rule and the pages on the right-hand side (consequent) **y** are recommended to the target user.

Sequential Discovery on the Weblog File

Sequential discovery in web usage mining is used to find past users' navigation patterns and predict the user's next visited page. In sequential discovery, transactions are viewed as sequences of pages that allow many useful and well-studied models to be used in discovering or analyzing user navigation patterns. One such model is the Markov model. This paper implements the Markov model, by considering the three parameters < A; S; T > where **A** is the set of all possible pages that can be visited by the user. **S** is the set of all possible states for which the Markov model is built, and **T** is the Transition Probability Matrix (TPM), where each entry $\mathbf{t}_{i,j}$ corresponds to the probability of performing the action **j** when the process is in state **i**.

Recommendation Layer

The recommendation layer is the last and the only online stage in the framework which is responsible to recommend or predict what kind of pages the user may prefer. This layer merges the aggregated usage profiles resulting from the clustering, frequent itemsets resulting from the association, and the prediction model resulting from sequential discovery techniques that are implemented in the pattern discovery layer to provide an effective and personalized recommendation to the user.

EXPERIMENTAL RESULTS

To properly test the effectiveness of the proposed framework discussed above the raw click-stream data collected from the access server log first must be cleaned and preprocessed into server sessions and then the different data mining techniques are applied to extract useful patterns to be used in generating recommendations to the users. This section applies the layers of the proposed framework to the collected data by implementing the pre-processing layer, the pattern discovery, and the recommendation layer to reach the paper's goal, which is providing recommendations based on users' interests that result in increasing customers (users) retention.

Data Set

The data has been collected from the access log file of the ftp://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html.The data contains 1727 records in Common Log Format (CLF) from 15 Jul 2010 to 22 Oct 2010. The size of the access log file was 100KB with 1727entries and 6 attributes.

Figure 8. A sample of raw web log data

The data preprocessing layer is the first initial step executed after collecting the data and it is the most complex and time-consuming layer.

Data Preprocessing Layer

Data preprocessing is the most important layer of web usage mining to clean web log data for discovering patterns. The log file is stored in a web server and may suffer from various data anomalies such data may be irrelevant, inconsistent, and noisy. The purpose of this layer is to remove any anomaly and to transform data into a structured format that enables to apply of the different data mining techniques. The anomalies were removed by applying the following as shown in figure 9.

Parsing Web Server Access Log File

Parsing the log's fields (Field extraction) is the process to separate the access log's fields from a single entry into multiple fields. As shown in Figure 8 data are stored in the access log file in a single field that needs to be separated the server uses different characters which work as separators, the most widely used separator characters are the space and (" "), so by parsing data, the data are transformed into fields and rows. The inputs of this step are the raw access log file. And the output is the parsed log file which contains fields like IP address, User ID, Date/Time, Request (method, URL, protocol), Status code, and bytes. Table 7 shows the data after applying the parsing step.

Significant Attributes Selection

After the Parsed access log file has been analyzed and the important attributes that provide valuable information about user behavior are obtained which are IP address, date/ time, request field (method/ URL/protocol), and status code. The input of this step is the Parsed access log file, and the output is the log file with only the important attributes. As shown in Table 8.

Data Selection

When a Web server receives an HTTP Request, it returns an HTTP response code to the client. This HTTP status code is three-digit number [10]. All records with status codes below 200 and above 299 are not used for analysis. Filtering of records is done by Microsoft excel to keep only the records

Figure 9. Data preprocessing implemented steps

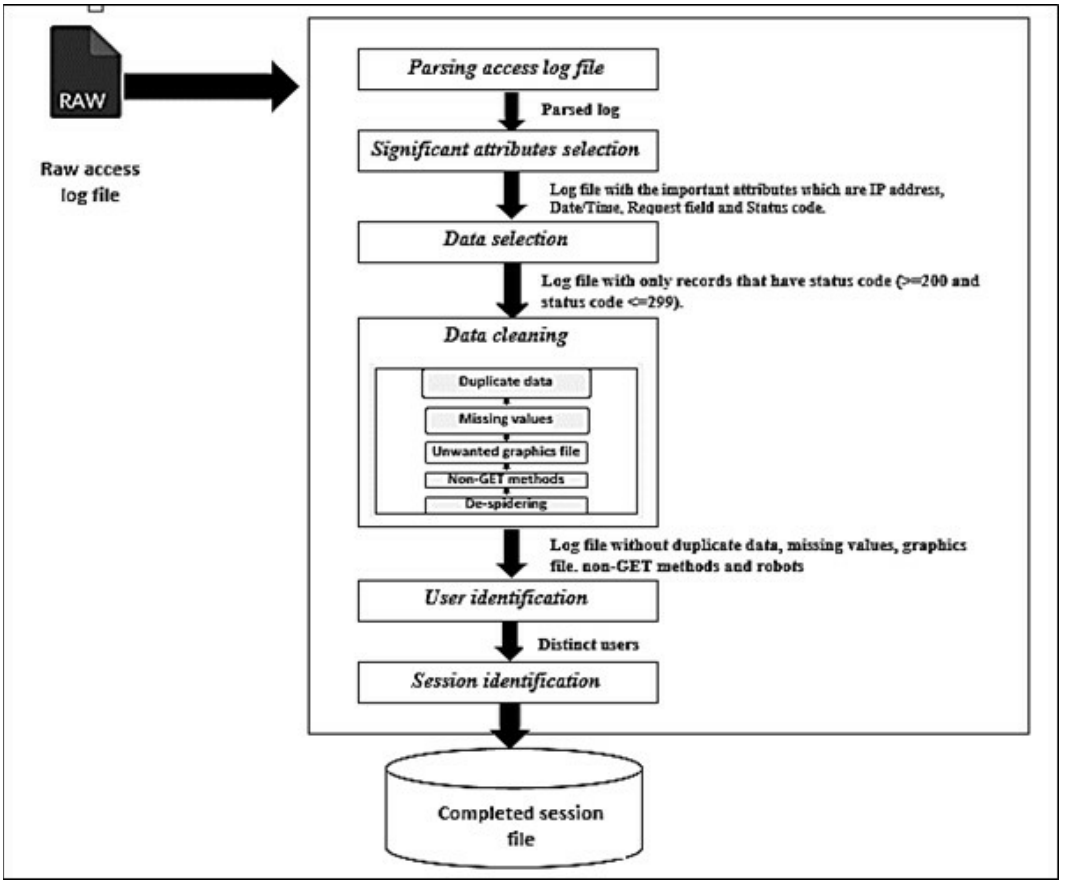


Table 7. Parsed web server access log file

IP address	User ID	Date/Time	Method	URL	Protocol	Status code	Bytes
10.1.109.54	-	[15/ Jul/2010:21:24:14]	GET	/displaytitle. php?id=10	HTTP/1.1	200	3024
10.1.11.81	-	[15/ Jul/2010:21:24:17]	GET	/displaytitle. php?id=10	HTTP/1.1	200	4029
10.1.11.81	-	[16/ Jul/2010:02:51:41]	GET	/mail/mailgraph.png	HTTP/1.1	200	2056
10.1.11.81	-	[16/ Jul/2010:03:00:21]	GET	/about-us/	HTTP/1.1	300	5486
10.1.11.81	-	[16/ Jul/2010:03:00:25]	GET	/release-schedule/	HTTP/1.1	404	3582
10.1.11.81	-	[16/ Jul/2010:03:00:29]	GET		HTTP/1.1	200	1256
10.1.114.150	-	[16/ Jul/2010:03:00:31]	GET	/faq/	HTTP/1.1	200	2478

Table 8. Significant Attributes of web server access log file

IP address	Date/Time	Method	URL	Protocol	Status code
10.1.109.54	[15/Jul/2010:21:24:14]	GET	/displaytitle.php?id=10	HTTP/1.1	200
10.1.11.81	[15/Jul/2010:21:24:17]	GET	/displaytitle.php?id=10	HTTP/1.1	200
10.1.11.81	[16/Jul/2010:02:51:41]	GET	/films/district-13/	HTTP/1.1	200
10.1.11.81	[16/Jul/2010:03:00:21]	GET	/news/	HTTP/1.1	200
10.1.11.82	[16/Jul/2010:02:00:25]	GET	/release-schedule/	HTTP/1.1	202
10.1.11.81	[16/Jul/2010:03:00:29]	GET	/trailers/	HTTP/1.1	200
10.1.114.151	[16/Jul/2010:04:00:31]	GET	/contact-us/	HTTP/1.1	200

with a status code (>=200 and status code <=299). The input of this step is the Parsed access log file with significance attributes and the output is the log file with only records that have status code (>=200 and status code <=299). By removing the unsuccessful status code, the number of records decreases from 1727 records to 1627 records, which means that 100 records containing unsuccessful status codes were removed.

Data Cleaning

Data cleaning is the step in the data pre-processing layer that is responsible for removing irrelevant and noise records that contain duplicated data, Missing values, unwanted graphics files, on-GET methods, and Robots. At the end of this step, all this noise will be eliminated as shown below.

1. Duplicated data

In the Microsoft Excel file, a range of cells was selected and the 'Remove Duplicates' tool was chosen to remove the duplicate records. By removing duplicated data, the number of records decreases from 1627 records to 1600 records, which means that 27 records contain duplicate data that were removed. As shown in the second and third- rows in Table 9.

2. Missing value

All the records that contain blank cells were removed. By removing records that contain missing values, the number of records decreases from 1600 records to 1595 records, which means that 5 records containing missing values were removed. As shown in the third row in Table 10.

Table 9. A sample of the access log file with duplicate records

IP address	Date/Time	Method	URL	Protocol	Status code
10.1.109.54	[15/ Jul/2010:21:24:14]	GET	/displaytitle.php?id=10	HTTP/1.1	200
10.1.11.81	[15/ Jul/2010:21:24:17]	GET	/displaytitle.php?id=10	HTTP/1.1	200
10.1.11.81	[15/ Jul/2010:21:24:17]	GET	/displaytitle.php?id=10	HTTP/1.1	200

Table 10. A sample of the access log file with Missing values

IP address	Date/Time	Method	URL	Protocol	Status code
10.1.109.54	[15/ Jul/2010:21:24:14]	GET	/displaytitle.php?id=10	HTTP/1.1	200
10.1.11.81	[16/ Jul/2010:02:51:41]	GET	/films/district-13/	HTTP/1.1	200
10.1.11.81	[16/ Jul/2010:03:00:21]	GET		HTTP/1.1	200
10.1.11.82	[16/ Jul/2010:02:00:25]	GET	/release-schedule/	HTTP/1.1	202

3. Unwanted graphics files

All records containing unwanted graphics files were removed. By removing these unwanted multimedia files, the number of records decreases from 1595 records to 1200 records, which means that 395 records containing unwanted graphics files were removed. As illustrated in the first and third records in Table 11.

4. Non-GET methods

As mentioned before, methods other than "GET" are removed because it does not reflect the user's interests and preferences. The filtering of records is done in Microsoft Excel by choosing the 'Filter' tool to keep records with only the 'GET' method. By removing non-GET methods, the number of records decreases from 1200 records to 1050 records, which means that 150 records contained methods other than "GET" were removed as shown in the last record in Table 12.

Table 11. A sample of the access log file with multi-media files

IP address	Date/Time	Method	URL	Protocol	Status code
10.1.102.54	[15/ Jul/2010:01:24:14]	GET	/mail/mailgraph.png	HTTP/1.1	200
10.1.11.81	[15/ Jul/2010:21:24:17]	GET	/displaytitle.php?id=10	HTTP/1.1	200
10.1.11.45	[16/ Jul/2010:02:51:41]	GET	/data/razor.jpg	HTTP/1.1	200

Table 12. A sample of the access log file with Non-GET methods

IP address	Date/Time	Method	URL	Protocol	Status code
10.1.11.82	[16/ Jul/2010:02:00:25]	GET	/release-schedule/	HTTP/1.1	202
10.1.11.81	[16/ Jul/2010:03:00:29]	GET	/trailers/	HTTP/1.1	200
10.1.114.151	[16/ Jul/2010:05:00:31]	POST	/cgi-bin	HTTP/1.1	200

5. De-Spidering

All records which are not related to humans were removed because the main goal is to understand users' behavior. By removing robots. The number of records decreases from 1050 records to 1000 records, which means that 50 records containing robots were removed.

As shown the irrelevant and noisy data were made up of a large percentage of the actual data size, so by cleaning this data, the number of records was significantly decreased, this indicates that the preprocessing of the web log file, helps to reduce and improve the quality of data. Which helps to increase the efficiency of patterns that will result from the pattern discovery layer.

User Identification

This step is responsible for identifying each distinct user. In this paper an e-commerce website is used, and all users are in a local network, so each user has a specific IP address without the agent's disturbance. 110 users were identified by assigning each user to a unique IP in the log file, which means that if a new IP address is identified this means that there is a new user. Each unique IP address gave an ID because it is easier to work with a single number like an ID in the pattern discovery layer instead of an IP address. As shown in Table 13.

Session Identification

The user session was identified by the Time gap method. A set of pages is considered single user session if those pages are requested within a specified period, which is 30 minutes, which is the default period used in many commercial applications. And if the request was given after that threshold, it is considered a new request (session). 186 users' sessions were identified, and session IDs were created based on this method. Table 14 illustrates a sample of the identified sessions.

The data pre-processing layer transforms and promotes the quality and accuracy of data that will be used as input for the pattern discovery layer. Figure 10 summarizes the data-preprocessing steps in detail.

Pattern Discovery Layer

In this layer, the pre-treated log file was analyzed to extract valuable patterns by using data mining techniques which are clustering, classification, association rule, and sequential discovery.

The main steps followed in this layer to extract patterns are as follow:

1. Choose the Data analytical tool to use.
2. Apply clustering over the weblog file to cluster users into groups of users with similar interests.
3. Apply classification over the log file to classify to which cluster (profile) the new user belongs.
4. Apply the Association rule to find correlations between web pages in data.
5. Apply sequential discovery to determine the next visited page.

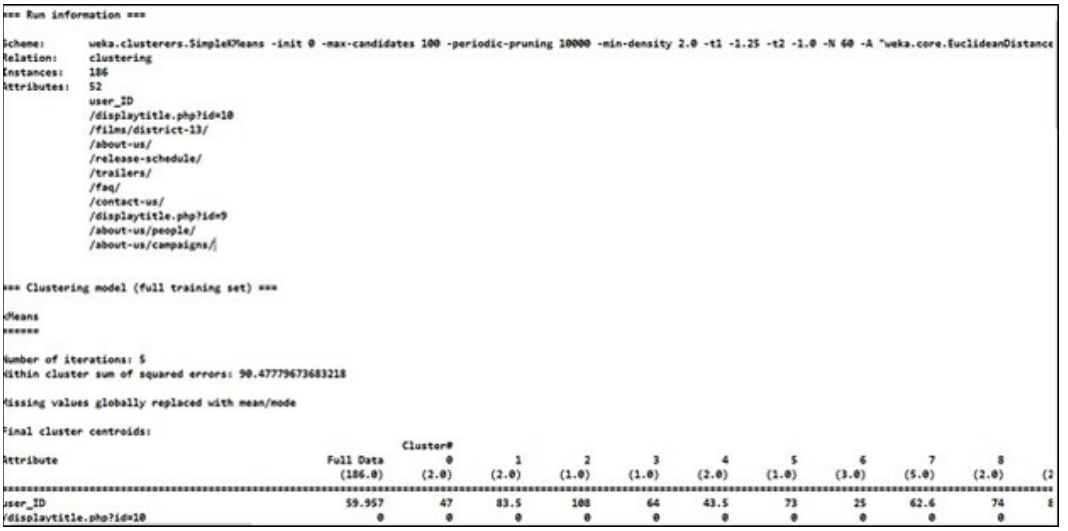
Table 13. A sample of unique users

USER ID	IP address	Date/Time	Method	URL	Protocol	Status code
1	10.1.109.54	[15/Jul/2010:21:24:14	GET	/displaytitle.php?id=10	HTTP/1.1	200
2	10.1.11.81	[15/Jul/2010:21:24:17	GET	/trailers/	HTTP/1.1	200
3	10.1.114.150	[16/Jul/2010:03:00:31	GET	/faq/	HTTP/1.1	200
4	10.1.12.172	[16/Jul/2010:03:00:35	GET	/contact-us/	HTTP/1.1	200

Table 14. A sample of identified users' sessions

USER ID	IP address	session ID	Date/Time	Method	URL	Protocol	Status code
1	10.1.109.54	1	[15/Jul/2010:21:24:14	GET	/displaytitle.php?id=10	HTTP/1.1	200
2	10.1.11.81	2	[15/Jul/2010:21:24:17	GET	/displaytitle.php?id=10	HTTP/1.1	200
	10.1.11.81	3	[16/Jul/2010:02:51:41	GET	/films/district-13/	HTTP/1.1	200
	10.1.11.81	4	[16/Jul/2010:03:00:21	GET	/about-us/	HTTP/1.1	200
	10.1.11.81		[16/Jul/2010:03:00:25	GET	/release-schedule/	HTTP/1.1	200
	10.1.11.81		[16/Jul/2010:03:00:29	GET	/trailers/	HTTP/1.1	200

Figure 10. K-means algorithm results



A Data Analytical Tool Used

Various data analysis tools can be used to mine data, such as Weka, Knime, and Rapidminer (RM). etc. Each of them has its advantages. The Weka analytical tool was used to perform mining on our data set. Mining data with the help of Weka resulted in useful information which can be used to solve problems such as who visited the website, build an intelligent website, and provide personalized recommendations.

Applying Clustering on a Weblog File

In this step, user transaction clustering was applied. User clustering was used to establish groups of users exhibiting similar browsing patterns. These clusters were used to create aggregated usage profiles, each profile is then represented as a vector in the original n-dimensional space. These aggregate profiles are used directly to provide recommendations to new users.

In this section, the clustering technique was applied to the preprocessed data in Table15 that come from the data pre-processing phase. The main goal of applying user clustering by the K-means algorithm is to analyze each segment of users and use these data to provide content to the users that are personalized to their interest through the recommendations offered to them. The resulting clusters were used in creating aggregate usage profiles which will be used in the recommendation layer to provide personalized recommendations to users.

In this paper, pages in each cluster are sorted according to their weights and lower weights pages that are below a specific threshold are filtered out and the remaining pages are used to form an "aggregate usage profile" which represents the interests and behavior of a significant group of users, as shown in Table 15. More formally, given a transaction cluster cl, we constructed the aggregate usage profile pr_{cl} by taking the centroid value of that cluster and used it to compute the weight of each page in that profile, the significance weight, weight (p, pr) of the page p within the aggregate profile pr_{cl} is given by equation 1:

$$\text{Weight(p,prcI)} = \frac{1}{|c1|} \sum_{s \in cI} W\ (\text{p,s})$$

- **| cl |** is the number of transactions in cluster cl.
- **W (p, s)** is the weight of page p in transaction vector s of cluster cl.
- The threshold **μ** is used to focus only on those pages in the cluster that appear in a sufficient number of vectors in that cluster which here is 0.5.

The pages '/about-us/' and **'**/release-schedule/' are the most significant pages characterizing the common interests of users in this profile. While pages as '/faq/', '/displaytitle.php?id=9' and '/news/?y=2009&m=8' are less significant. When a new user enters the website, he/she must be classified to belong to any of the aggregate profiles, to provide him/her with recommendations. This is done by analyzing his navigational path and determining the profile that they belong to. The classification technique is used to classify the new user (active session) to the right profile based on his/her neighbors.

Applying Classification on a Weblog File

K-nearest neighbor (KNN) was used to measure the correlation between the current user profile (active session) and the aggregated profiles resulting from the clustering technique to find a profile in the database similar to the preferences or characteristics of the target user (Nikam & technology, 2015)

Applying Association Rule on a Weblog File

The association rule methods such as the A priori algorithm were used to find groups of pages occurring frequently together in many transactions. Such groups of pages are referred to as frequent itemsets.

After applying the apriori algorithm, the resulting frequent itemsets are stored in an ascending order to be used later in the recommendation layer to provide a recommendation set to the target user. Weka tool provides us with the frequent itemsets, as shown in Table 16. When the target user enters

Table 15. A sample of aggregated profiles

Aggregated usage profile 0	
Page	**Weight**
/about-us/	1
/release-schedule/	1
/faq/	0.5
/displaytitle.php?id=9	0.5
/news/?y=2009&m=8	0.5

Table 16. A sample of frequent itemsets

State	Next visit	Probability
/Faq/	/trailers/	0.5
/contact-us/	/about-us/	0.4
/displaytitle.php?id=9	/release-schedule/	0.5
/about-us/people/	/about-us/campaigns/	0.6
/about-us/campaigns/	/about-us/case-studies/blood/	0.3

the website his/her preferences are matched against the antecedent **x** of each rule and the pages on the right-hand side (consequent) **y** of the matching rules are recommended to the target user.

Applying Sequential Discovery on a Weblog File

In sequential discovery, web transactions are viewed as sequences of pages that allow many useful models to be used in discovering or analyzing user navigation patterns. One such approach is the Markov model. After applying the Markov model of the data, the results were stored to make a recommendation to the target user with the page that the user may visit in the future. As shown in Table 17.

The clustering, association rules, and sequential discovery techniques have some important factors that affect the quality of their results that must be taken into account while providing recommendations which are:

- In clustering, the change in the number of clusters, results in a different recommendation set to the same user and results in making the KNN each time attach the user to belong to a different profile.
- In the association rule, the values of Support and confidence affect the generated rules which means that some of the significant rules may be eliminated due to the incorrect values of support or confidence.
- In sequential discovery, the results of the first-order Markov model have low accuracy and the higher Markov model suffers from noise.

Table 17. A sample of the predicted page for each state

Large Itemsets L(2): /displaytitle.php?id=10=T /films/district-13/=T 3 /films/district-13/=T /about-us/=T 2 /about-us/=T /trailers/=T 17 /about-us/people/=T /about-us/campaigns/=T 11 /about-us/people/=T /displaytitle.php?id=8=T 6
Large Itemsets L(3): /displaytitle.php?id=10=T /about-us/=T /release-schedule/=T 6 /about-us/=T /release-schedule/=T /about-us/campaigns/=T 8 /trailers/=T /displaytitle.php?id=13=T /displaytitle.php?id=16=T 2 /faq/=T /contact-us/=T /displaytitle.php?id=9=T 5 /displaytitle.php?id=9=T /displaytitle.php?id=8=T /displaytitle.php?id=12=T 2
Large Itemsets L(4): /displaytitle.php?id=9=T /about-us/campaigns/=T /displaytitle.php?id=8=T /displaytitle.php?id=13=T 2 /about-us/people/=T /about-us/campaigns/=T /displaytitle.php?id=8=T /about-us/case-studies/blood/=T 2 /faq/=T /about-us/people/=T /about-us/campaigns/=T /displaytitle.php?id=8=T 4

Therefore, our proposed framework integrated clustering, classification, association rule, and sequential discovery to overcome the imperfection of each technique and provide the user with recommendations based on:

- His/Her similarity with the available profiles.
- The relationship between the pages exists in his session.
- The predicted next visited page.

Recommendation Layer

The recommendation layer is the last and the only online stage in the framework which is responsible to recommend or predict what kind of pages the user may prefer. This layer merges the aggregated usage profiles, frequent itemsets, and the prediction model resulting from the pattern discovery layer to provide an effective and personalized recommendation to the user.

Recommendation Based on Clustering and Classification

Given a target user whose User ID is 103 and who has accessed a set of pages. The KNN classified the target user to the most similar aggregate usage profile, as shown in Table 18 and Table 19.

The results of the classification algorithm predict that profile11 is the most similar profile to user 103 active session. Based on the similarity and the weights of each page within the profile a recommendation score was computed that leads to results between 0 and 1. If the page is in the current active session, then its recommended value is set to 0. The pages are recommended to the user based on the recommendation score so pages are sorted in descending order and pages with the highest recommendation score are the most recommended pages. Table 20 shows the recommended pages.

After obtaining the recommendation set, all the pages whose recommendation score satisfies a minimum recommendation threshold (ρ=0.4) were recommended to the target user, which in this case are '/displaytitle.php?id=9' with a recommendation score of 0.7 and '/news/?y=2009&m=8' with also recommendation score 0.7 and the other pages is filtered out. The pages are recommended to the user based on the recommendation score so pages are sorted in descending order and pages with the highest recommendation score are the most recommended pages as represented in equation 2:

$$\mathrm{Re}\,c\left(S,p\right)=\sqrt{weight\left(p,C\right)}.match\left(S,C\right)$$

Table 18. Aggregate usage profile11

Profile11	Page	/about-us/	/release-schedule/	/trailers/	/Faq/	/displaytitle.php?id=9	/news/	/news/?y=2009&m=8
	Weight	1	1	0.5	0.5	0.5	1	0.5

Table 19. Active user session

User 103 (active session)	Page				
	/about-us/	/release-schedule/	/trailers/	/Faq/	/news/

Table 20. Clustering-Recommendation set

Page	Recommendation Score
/displaytitle.php?id=9	0.7
/news/?y=2009&m=8	0.7

Recommendation Based on Association Mining

After that, the results of association rule mining are used to produce a recommendation for the target user. The top-ranked pages in the consequent whose confidence is greater than or equal to a specified threshold are recommended to the target user.

Firstly the most match frequent itemset was found in the current user session to find pages to recommend for the target user. For instance, given the active session of user 103 and a group of frequent itemsets that resulted from the pattern discovery layer, all the frequent itemsets of size equal to, (session size +1) containing the current session were considered. In this case, all the frequent itemset of size equal to (session size +1) was searched. As shown in Table 21.

After selecting the sets that match antecedents, the confidence for each of them is calculated by dividing the support of the rule by the support of the sliding window and after obtaining the confidence of each set, a minimum confidence threshold (ρ=0.4) is specified which means that any rules with confidence less than 0.4 are filtered out and the remaining rules are used to make a recommendation to the target user, as shown in Table 22.

Recommendation Based on Sequential Discovery

The last step in this layer is to make recommendations depending on the sequence of pages in the target user session. The result of the pattern discovery layer is used to predict the user's next visit to make a recommendation based on this prediction. As shown in Table 23.

Table 21. A sample of the resulted frequent itemset

/films/district-13/=T /contact-us/=T /about-us/people/=T /about-us/campaigns/=T /displaytitle.php?id=8=T 2
/about-us/=T /release-schedule/=T /trailers/=T /faq/=T /contact-us/=T 7
/films/district-13/=T /faq/=T /about-us/people/=T /about-us/campaigns/=T /displaytitle.php?id=8=T 2
/about-us/=T /release-schedule/=T /trailers/=T /faq/=T /displaytitle.php?id=9=T 4
/about-us/=T /release-schedule/=T /trailers/=T /faq/=T /about-us/people/=T 10
/displaytitle.php?id=10=T /displaytitle.php?id=9=T /about-us/people/=T /about-us/campaigns/=T /displaytitle.php?id=8=T 2
/about-us/=T /release-schedule/=T /trailers/=T /faq/=T /about-us/campaigns/=T 7
/about-us/=T /release-schedule/=T /trailers/=T /faq/=T /displaytitle.php?id=8=T 5
/about-us/=T /release-schedule/=T /trailers/=T /faq/=T /displaytitle.php?id=11=T 2
/about-us/=T /release-schedule/=T /trailers/=T /faq/=T /about-us/case-studies/blood/=T 2
/about-us/=T /release-schedule/=T /trailers/=T /faq/=T /displaytitle.php?id=13=T 2
/about-us/=T /release-schedule/=T /trailers/=T /contact-us/=T /displaytitle.php?id=9=T 4

Table 22. Association- Recommendation set

Page	Recommendation Score
/about-us/people/	0.9
/contact-us/	0.6
/about-us/campaigns/	0.6
/displaytitle.php?id=8	0.5

Table 23. Sequential discovery-Recommendation set

Page	Recommendation Score
/displaytitle.php?id=10	0.8

Recommendation Based on Merging Techniques

As illustrated in Table 24, each of the mentioned techniques resulted in a recommendation set that differs from one another, despite the existence of the same target user with the same user session and with a fixed recommendation threshold in the three techniques.

As a result, a merged recommendation set was created from the three techniques and offered to the target user which will result in more accurate and comprehensive recommendations offered to the user and overcome the imperfection in the recommendation set supplied by each technique. This recommendation set is added to the last visited page in the user active session and pages are sorted according to their recommendation score, as shown in Table 25.

Performance Evaluation

To test and evaluate the benefits of using our proposed framework for creating personalized recommendations for users. To prove whether it is enhancing customer retention. The accuracy of recommendations offered to customers based on the clustering and classification technique, the association rule, and the sequential discovery technique is compared against the merging technique.

Each Clustering and classification, association rule, and sequential discovery technique is used individually to retrieve the suitable recommendation set according to the needs and preferences of each customer. And then the integrated (merged) technique is used to retrieve the recommendation

Table 24. Recommendation set of each technique

Clustering and Classification		Association mining		Sequential discovery	
Page	score	Page	score	Page	score
/displaytitle.php?id=9	0.7	/about-us/people/	0.9	/displaytitle.php?id=10	0.8
/news/?y=2009&m=8	0.7	/contact-us/	0.6		
		/about-us/campaigns/	0.6		
		/displaytitle.php?id=8	0.5		

Table 25. Merged Recommendation set

Page	Recommendation Score
/about-us/people/	0.9
/displaytitle.php?id=10	0.8
/displaytitle.php?id=9	0.7
/news/?y=2009&m=8	0.7
/contact-us/	0.6
/about-us/campaigns/	0.6
/displaytitle.php?id=8	0.5

set R to the customer the effectiveness of a proposed framework is measured in terms of accuracy (precision, coverage, and F-measure values).

Precision is the ratio of the number of retrieved pages that are relevant to the total number of retrieved pages based on the preferences and interests of each customer as represented in equation 3. Precision measures the probability that a recommended page is relevant. Coverage is the ratio of the number of retrieved pages that are relevant to the total number of relevant pages as represented in equation 4. Coverage measures the probability that a relevant page is recommended. Precision and coverage are calculated using the following equations (Jain, Seeja, & Jindal, 2021):

$$\text{Precision (R,t)} = \frac{\left|R \cap \left(t - w\right)\right|}{\left|R\right|}$$

$$\text{Coverage(R,t)} = \frac{\left|R \cap \left(t - w\right)\right|}{\left|t - w\right|}$$

Neither of these measures individually is sufficient to evaluate the performance of the recommender system, however, they are both critical. Because a low precision in this context will likely result in angry customers who are not interested in the recommended pages, while low Coverage will result in the inability of the recommender system to produce relevant recommendations at critical points in the user's interaction with the site. Both negative phenomena have led us to use a single measure that captures high precision and high coverage which is the F1 measure which is represented in equation 5. The F1 attains its maximum value when both precision and coverage are maximized (Berbague et al., 2021).

The result of testing each technique individually and the merged technique is illustrated in Table 26:

$$\text{F1(R,1)} = \frac{2 * Precision\left(R,1\right) * Coverage\left(R,1\right)}{Precision\left(R,1\right) + Coverage\left(R,1\right)}$$

The result of testing the proposed framework illustrated that the recommendations based on the merged technique achieve a strong accuracy with a precision value of 74%, the coverage value of 100%, and the average overall efficiency of the F-measure was 86%. The association technique outperformed the other techniques in terms of precision while showing poor coverage concerning clustering and classification technique and merged technique. While merged technique attained a much higher overall coverage, which led to a relatively good F1 score.

Table 26. Summary of the precision, coverage, and F1 for each technique

Technique	Evaluation measures		
	Precision	**Coverage**	**F1**
Clustering and Classification	0.77 (77%)	0.6 (60%)	0.67 (67%)
Association rule	0.83 (83%)	0.52 (52%)	0.64 (64%)
Sequential discovery	0.76 (76%)	0.24 (24%)	0.35 (35%)
Merged technique	0.74 (74%)	1 (100%)	0.86 (86%)

CONCLUSION AND FUTURE WORK

Limitations: Lack of customer satisfaction in the highly competitive market.

Practical implications: The effectiveness of the proposed framework was tested and evaluated. The results of the test proved that the proposed a framework achieves its objectives that resulting in increased customer retention. The framework achieves strong and accurate results in making precise recommendations for customers with a precision value of 74%, a coverage value of 100%, and the average overall efficiency of the F-measure was 86%.

Managerial implications: This paper aims at developing personalized recommendations to users of commercial websites to increase their satisfaction and support organizations to retain their customers. The proposed framework empirically demonstrated the feasibility of understanding customers' behaviors through tracking their history on the commercial website and providing high-quality and accurate recommendations that match their preferences and interests by merging clustering, classification, association rule, and sequential discovery techniques.

Future research: Direction applying deep learning techniques to maximize the percentage of accuracy.

FUNDING STATEMENT

No funding was received for conducting this study.

DATA AVAILABILITY STATEMENT

Data is contained within the article.

CONFLICTS OF INTEREST

The author declare that she has no conflicts of interest to report regarding the present study.

REFERENCES

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. doi:10.1016/j.aci.2014.10.001

Aldekhail, M. (2016). Application and significance of web usage mining in the 21st century: A literature review. *International Journal of Computer Theory and Engineering*, *8*(1), 41–47. doi:10.7763/IJCTE.2016.V8.1017

Bahari, T. F., & Elayidom, M. S. J. P. c. s. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. *46*, 725-731.

Bandyopadhyay, S., Thakur, S., & Mandal, J. (2017). *Product recommendation for E-commerce data using association rule and apriori algorithm.* Paper presented at the International Conference on Modelling and Simulation. Semantic Scholar.

Bayir, M. A., & Toroslu, I. H. (2022). Maximal paths recipe for constructing Web user sessions. *World Wide Web (Bussum)*, *25*(6), 1–31. doi:10.1007/s11280-022-01024-3

Berbague, C. E., Karabadji, N. E.-i., Seridi, H., Symeonidis, P., Manolopoulos, Y., & Dhifli, W. (2021). An overlapping clustering approach for precision, diversity and novelty-aware recommendations. *Expert Systems with Applications*, *177*, 114917. doi:10.1016/j.eswa.2021.114917

Britvin, A., Alrawashdeh, J. H., & Tkachuk, R. (2022). Client-Server System for Parsing Data from Web Pages. *Advances in Cyber-Physical Systems*, *7*(1), 8–13. doi:10.23939/acps2022.01.008

Cai, L., Wang, H., Jiang, F., Zhang, Y., & Peng, Y. (2022). A new clustering mining algorithm for multi-source imbalanced location data. *Information Sciences*, *584*, 50–64. doi:10.1016/j.ins.2021.10.029

Chavda, S., Jain, S., Panchal, N., & Valera, M. (2017a). Recent trends and novel approaches in web usage mining. [IRJET]. *Int. Res. J. Eng. Technol.*, *4*, 1319–1322.

Dogan, O., Kem, F. C., & Oztaysi, B. (2022). Fuzzy association rule mining approach to identify e-commerce product association considering sales amount. *Complex & Intelligent Systems*, *8*(2), 1551–1560. doi:10.1007/s40747-021-00607-3

Elhebir, M. H. A., Abraham, A. J. J. N., & Computing, I. (2015). *Discovering Web Server Logs Patterns Using Clustering and Association Rules Mining.*, *3*, 159–167.

Ganibardi, A., & Ali, C. A. (2018). *Web Usage Data Cleaning.* Paper presented at the International Conference on Big Data Analytics and Knowledge Discovery. Springer. doi:10.1007/978-3-319-98539-8_15

Hao, S., Zhaoxiang, S., & Bingbing, Z. (2017). *A user clustering algorithm on web usage mining.* Paper presented at the 2017 First International Conference on Electronics Instrumentation & Information Systems (EIIS). IEEE. doi:10.1109/EIIS.2017.8298745

Hariharakrishnan, J., Mohanavalli, S., & Kumar, K. S. (2017). *Survey of pre-processing techniques for mining big data.* Paper presented at the 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP). IEEE. doi:10.1109/ICCCSP.2017.7944072

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal, 16*(3), 261-273.

Ismail, M., Ibrahim, M. M., Sanusi, Z. M., Nat, M. J. I. J. (2015). Data Mining in electronic commerce: benefits and challenges. *Int. J. Communications, Network and System Sciences*, *8*(12), 501.

Jain, S., Seeja, K., & Jindal, R. (2021). A fuzzy ontology framework in information retrieval using semantic query expansion. *International Journal of Information Management Data Insights*, *1*(1), 100009. doi:10.1016/j.jjimei.2021.100009

Jardine, E. (2021). Policing the cybercrime script of darknet drug markets: Methods of effective law enforcement intervention. *American Journal of Criminal Justice*, *46*(6), 980–1005. doi:10.1007/s12103-021-09656-3

Kaadoud, I. C., Rougier, N. P., & Alexandre, F. (2022). Knowledge extraction from the learning of sequences in a long short term memory (LSTM) architecture. *Knowledge-Based Systems*, *235*, 107657. doi:10.1016/j.knosys.2021.107657

Kaur, J., & Garg, K. (2019). Efficient Management of Web Data by Applying Web Mining Pre-processing Methodologies [Springer.]. *Software Engineering*, *731*, 115–122. doi:10.1007/978-981-10-8848-3_11

Kaur, N., & Aggarwal, H. (2015). Web log analysis for identifying the number of visitors and their behavior to enhance the accessibility and usability of website. *International Journal of Computers and Applications*, *110*(4).

Kumar, A., Bhushan, B., Pokhriya, N., Chaganti, R., & Nand, P. (2022). *Web Mining and Web Usage Mining for Various Human-Driven Applications Advanced Practical Approaches to Web Mining Techniques and Application*. IGI Global.

Kumar, V., & Ayodeji, O. G. (2021). E-retail factors for customer activation and retention: An empirical study from Indian e-commerce customers. *Journal of Retailing and Consumer Services*, *59*, 102399. doi:10.1016/j.jretconser.2020.102399

Lopes, P., & Roy, B. (2015). Dynamic recommendation system using web usage mining for e-commerce users. *Procedia Computer Science*, *45*, 60–69. doi:10.1016/j.procs.2015.03.086

Mehra, J., & Thakur, R. (2018). An effective method for web log preprocessing and page access frequency using web usage mining. *International Journal of Applied Engineering Research: IJAER*, *13*(2), 1227–1232.

Mittal, R., Malik, V., Rattan, V., & Jhamb, D. (2021). *Performance comparison of tree-based machine learning classifiers for web usage mining.* Paper presented at the Proceedings of International Conference on Communication, Circuits, and Systems. Springer. doi:10.1007/978-981-33-4866-0_47

Moreno, M. N., Segrera, S., López, V. F., Muñoz, M. D., & Sánchez, Á. L. (2016). Web mining based framework for solving usual problems in recommender systems. A case study for movies′ recommendation. *Neurocomputing*, *176*, 72–80. doi:10.1016/j.neucom.2014.10.097

Neelima, G., & Rodda, S. (2016). *Predicting user behavior through sessions using the web log mining.* Paper presented at the 2016 International Conference on Advances in Human Machine Interaction (HMI). IEEE. doi:10.1109/HMI.2016.7449167

Nguyen, M.-T., Diep, T.-D., Hoang Vinh, T., Nakajima, T., & Thoai, N. (2018). *Analyzing and visualizing web server access log file.* Paper presented at the International Conference on Future Data and Security Engineering. Springer. doi:10.1007/978-3-030-03192-3_27

Nikam, S. S. J. (2015). A comparative study of classification techniques in data mining algorithms. *International research journal of Computer Science and Technology*, *8*(1), 13-19.

Panigrahi, R., Borah, S., Bhoi, A. K., Ijaz, M. F., Pramanik, M., Jhaveri, R. H., & Chowdhary, C. L. (2021). Performance assessment of supervised classifiers for designing intrusion detection systems: A comprehensive review and recommendations for future research. *Mathematics*, *9*(6), 690. doi:10.3390/math9060690

Patel, K. D., & Parikh, S. M. J. I. J. C. A. (2017).. . *Preprocessing on Web Server Log Data for Web Usage Pattern Discovery.*, *165*(10), 29–32.

Roy, R. (2021). *Predicting User's Web Navigation behaviour using AMD and HMM Approaches.* Paper presented at the IOP Conference Series: Materials Science and Engineering. doi:10.1088/1757-899X/1074/1/012031

Saleh Ibrahim, Y., Muhammed, Y., Al-Douri, A. T., Faisal, M. S., Mohamad, A. A. H., Al-Husban, A., & Birhan, M. (2022). Discovery of Knowledge in the Incidence of a Type of Lung Cancer for Patients through Data Mining Models. *Computational Intelligence and Neuroscience*, *2022*, 2022. doi:10.1155/2022/6058213 PMID:35685154

Samboteng, L., Rulinawaty, R., Kasmad, M. R., Basit, M., & Rahim, R. (2022). Market Basket Analysis of Administrative Patterns Data of Consumer Purchases Using Data Mining Technology. *Journal of Applied Engineering Science*, *20*(2), 339–345. doi:10.5937/jaes0-32019

Sathya, M., & Devi, P. I. (2017). *Apriori algorithm on web logs for mining frequent link.* Paper presented at the International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). IEEE. doi:10.1109/ITCOSP.2017.8303127

Sellamy, K., Fakhri, Y., Boulaknadel, S., Moumen, A., Hafed, K., Jamil, H., & Lakhrissi, Y. (2018). *Web mining techniques and applications: Literature review and a proposal approach to improve performance of employment for young graduate in Morocco.* Paper presented at the 2018 International Conference on Intelligent Systems and Computer Vision (ISCV). IEEE. doi:10.1109/ISACV.2018.8354043

Selvy, P. T., Anitha, M. M., Varthan, L. V., Sethupathi, P., & Adharsh, S. (2022). Intelligent Web Data Extraction System for E-commerce. *Journal of Algebraic Statistics*, *13*(3), 63–68.

Sengottuvelan, P., Lokeshkumar, R., & Gopalakrishnan, T. (2017). An improved session identification approach in web log mining for web personalization. *Journal of Internet Technology*, *18*(4), 723–730.

Sengottuvelan, P., Lokeshkumar, R., & Gopalakrishnan, T. J. 網. (2015). An Improved Session Identification Approach in Web Log Mining for Web Personalization.

Shivaprasad, G., Reddy, N. S., & Acharya, U. D. (2015). Knowledge Discovery from Web Usage Data: An Efficient Implementation of Web Log Preprocessing Techniques. *International Journal of Computer Application, 111*(13).

Singh, M. (2020). Scalability and sparsity issues in recommender datasets: A survey. *Knowledge and Information Systems*, *62*(1), 1–43. doi:10.1007/s10115-018-1254-2

Sirichanya, C., & Kraisak, K. (2021). Semantic data mining in the information age: A systematic review. *International Journal of Intelligent Systems*, *36*(8), 3880–3916. doi:10.1002/int.22443

Soltani, Z., & Navimipour, N. J. (2016). Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research. *Computers in Human Behavior*, *61*, 667–688. doi:10.1016/j.chb.2016.03.008

Srinivas, A. (2017). A survey on preprocessing of web-log data in web usage mining. *International Journal for Modern Trends in Science and Technology, 3*.

Subramaniyaswamy, V., & Logesh, R. (2017). Adaptive KNN based recommender system through mining of user preferences. *Wireless Personal Communications*, *97*(2), 2229–2247. doi:10.1007/s11277-017-4605-5

Suchacka, G., & Chodak, G. (2017). Using association rules to assess purchase probability in online stores. *Information Systems and e-Business Management*, *15*(3), 751–780. doi:10.1007/s10257-016-0329-4

Swamy, K. R., Babu, G. H., Venkatasubbaiah, R. (2015). Identification of Frequent Item Search Patterns Using APRIORI Algorithm and WEKA Tool. *International Journal of Innovative Technology and Research, 3*(5), 2401-2403.

Thorat, P. B., Goudar, R. M., & Barve, S. (2015). Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computers and Applications*, *110*(4), 31–36. doi:10.5120/19308-0760

Tiwari, S., Gupta, R. K., & Kashyap, R. (2019). *To enhance web response time using agglomerative clustering technique for web navigation recommendation Computational Intelligence in Data Mining*. Springer.

Zhang, J., Lin, Y., Lin, M., & Liu, J. J. A. I. (2016). An effective collaborative filtering algorithm based on user preference clustering. *45*(2), 230-240.

Zhu, H. (2022). *Constructing an Enterprise Management System Using an Improved Clustering Algorithm*. Mobile Information Systems. doi:10.1155/2022/9119452

Zidan, K., Sbahi, S., Hejjaj, A., Ouazzani, N., Assabbane, A., & Mandi, L. (2022). Removal of bacterial indicators in on-site two-stage multi-soil-layering plant under arid climate (Morocco): prediction of total coliform content using K-nearest neighbor algorithm. *Environmental Science and Pollution Research*, 1-14. .

Shimaa Ouf is currently an Assistant Professor in the Business Information Systems Department, Faculty of Commerce and Business Administration, Helwan University. She was born in Cairo, Egypt. She received a Diploma degree in Business Information Technology from the Faculty of Computers and Information, Helwan University, Egypt, Excellent and ranked first in a class of 2010. M.Sc. degree in Information Systems from Faculty of Computers and Information, Helwan University, Egypt, 2012, Thesis Topic: An Enhanced E-Learning Ecosystem Based on Integration between Cloud Computing and Web 2.0. Ph.D. degree in Information System from Faculty of Computers and Information, Helwan University, Egypt, 2017, Dissertation Topic: A proposed Paradigm for Smart E-Learning Ecosystem Based on Semantic Web. She has published many articles in international journals and conferences in E-learning Ecosystem, Web 2.0 Technologies, Cloud Computing, Smart learning environment, Personalized E-learning Ecosystem using Ontology and Semantic Web Rule Language, Business Intelligence, Big Data, and Blockchain. She has deep experience in personalizing the learning environment using semantic web technologies. She is an active reviewer for numerous international journals. Prof/Yehia Helmy, Merna Ashraf.