A Novel Web Recommendation Model Based on the Web Usage Mining Technique

Dalia L. Elsheweikh

Department of Computer Science, Faculty of Specific Education, Mansoura University, Mansoura, Egypt
Email: dalia_lotfy@mans.edu.eg

***Abstract*—Most models of automated web recommender systems depend on data mining algorithms to discover useful navigational patterns from the user's previous browsing history. This paper presents a new model for developing a collaborative web recommendation system using a new technique for knowledge extraction. The proposed model introduces two techniques: cluster similarity-based technique and rule extraction technique to provide proper recommendations that meet the user's needs. A cluster similarity-based technique groups the sessions that share common interests and behaviors according to a new similarity measure between the web users' sessions. The rule extraction technique, which is based on a trained Artificial Neural Network (ANN) using a Genetic Algorithm (GA), is performed to discover groups of accurate and comprehensible rules from the clustering sessions. For extracting rules that belong to a specific cluster, GA can be applied to get the perfect values of the pages that maximize the output function of this cluster. A set of pruning schemes is proposed to decrease the size of the rule set and remove non-interesting rules. The resulting set of web pages recommended for a specific cluster is the dominant page in all rules that belong to this cluster. The experimental results indicate the proposed model's efficiency in improving the classification's precision and recall.**

***Keywords*—collaborative web recommender systems, web usage mining, web log file, Artificial Neural Network (ANN), Neural Network (NN), Genetic Algorithm (GA), clustering technique, knowledge extraction technique**

I. Introduction

There is a huge increase in the amount of data accessible on the World Wide Web (WWW). Nowadays, online browsers allow simple access to a plethora of various kinds of data sources. Search engines index huge numbers of pages, up to the millions, making it difficult to discover the information needed. The redundancy of resources encouraged the demand to improve automatic mining techniques in the WWW, which led to the emergence of the term "web mining" [1]. Web mining is a nascent field of research for detecting useful and interesting patterns from the data relating to websites. Web mining is formed from three fields based on the nature of the data: web content mining, web structure data mining, and web usage mining [2]. Web content mining is aimed at getting beneficial information from the vast amount of web data. These data include various types of data, such as audio, video, symbolic, metadata, and hyperlinked data [3]. Web structure mining is an instrument utilized for recognizing the correlation between web pages related to information or a direct link connection. This structure data can be identified by providing a web structure schema using web page database techniques. Structure mining lets search engines determine and return a search query straight to the linking web page from the website where it existed. Structure mining is widely utilized for extracting formerly unknown relationships among web pages [4]. Web Usage Mining (WUM) analyses the usage patterns of web sites so as to gain a grasp of the interests and behaviors of web users. It should be noted that content mining and structure mining profit from actual or major data on the web, while usage mining mines minor data produced by the users' reactions to the web [5]. Web usage data contains detailed information like web server access logs, proxy server logs, referrer logs, browser logs, error logs, user profiles, registration data, cookies, user sessions, user searches, bookmark folders, and any other data obtained through the users' interaction with the web [6]. Web usage mining techniques have been applied to many practical uses, such as web personalization system development, website modification, and e-business intelligence [7]. Web personalization can be identified as the operation of harmonizing web pages using a user's browsing patterns and interests. With the huge number of websites and web pages found on the internet, guiding users to the web pages of their respective fields of concern has become a complex task. For years, varied approaches have been proposed and developed, and each approach has been adopted for producing personalized web recommendations [8]. The works contributed to the personalization of a website are based on user profiles and web usage patterns. Clustering of the user profiles may be used to find groups of users sharing similar features and interests, but its success depends on valid user feedback. This approach may be useful for web sites such as online banking, insurance companies, and corporate and business houses that maintain authentic users and sometimes provide relevant returns to the users based on their categories [9]. Access logs can be used to obtain access patterns that improve web access performance [10]. System improvement occurs through web traffic behavior. Web usage mining of

Manuscript received March 28, 2023; revised May 16, 2023; accepted June 6, 2023; published October 8, 2023.

patterns produces the way to conception web traffic behavior that can be utilized to transact with policies on web caching, network transmission, load balancing, or data distribution [11]. Site amendment is the process of amending the website and increasing the fineness of the design and components to find out the users' interests. Therefore, web usage mining can be utilized to supply the organization with information about its website so as to improve user browsing activities [12]. Web usage mining can be used as a way for business intelligence aims to get hidden patterns and business strategies about its customers and web data. The knowledge acquired from web usage patterns can be directly used to increase efficacy in managing activities related to e-business, e-services, and e-education. Accurate web usage information can assist in winning new clients, keeping existing clients, enhancing cross-marketing and sales, increasing the effectiveness of advertising campaigns, following clients' leaving, and so on. Information usage can be applied to improve web server performance by improving appropriate perfecting and caching strategies for decreasing the server response time. User profiles could be created by merging the navigation paths of users with other data features such as page viewing time, hyperlink structure, and page content [13]. There are many techniques that have been utilized in web usage mining for discovering rules and patterns, such as statistical analysis [14], association rules [15], sequential patterns [16], clustering [17], classification [18], and dependency modeling [19], which are applied to web server logs.

Hence, the primary goal of this work is to recommend the appropriate pages based on patterns discovered from previous users' sessions. A novel rule extraction technique can be founded on a trained neural network using a genetic algorithm to extract a group of comprehensive and interesting rules from the data of the users' sessions. This work is constructed as follows: Section II presents web usage mining. Section III introduces the general proposed model and its phases. The proposed recommendation system is introduced in Section IV. Section V presents the experiment and analyses its results. At last, Section VI provides the study's conclusion.

II. Web Usage Mining

Web Usage Mining (WUM) is crucial to improving the effectiveness of website management, building adaptive websites, business and support services, personalization, network traffic flow analysis, etc. Web usage mining tried to detect interesting and helpful patterns from the minor data acquired based on users' interactions and the web [20]. WUM is composed of three stages: the pre-processing stage, the pattern discovery stage, and the pattern analysis stage. Web servers are rated as the most popular source of data of all kinds. Often, the web server records the access activities of web site users in web server logs, such as IP address, date and time, method, URL, referrer, protocol, HTTP errors, number of bytes transferred, and agent [21]. The other job is the treatment of outliers, errors, and missing data, which can happen because of the inherent in web browsing. Generally, data pre-processing involves data cleaning, user identification, and user session identification [22].

Data cleaning aims at separating unrelated and refined log entries for the mining process. The records that have filename suffixes of GIF, JPEG, CSS, etc. that may be downloaded without a direct request from the user must be removed. Other removal records include HTTP errors, records created by crawlers, etc., that cannot really show the user's behavior [23]. The records with status codes less than 200 or more than 299 are deleted.

User identification is a very important part of the server session identification process. After the log file has been cleaned, we need to identify users. There are different approaches to user identification. The first approach is to use the user registration data. For websites that ask for user registration (username and password), the user identification task is straightforward and guaranteed to be correct. Thus, the log file also contains the user login, which is utilized for user identification. However, not all sites can perform this task because it makes the users afraid and is inappropriate for web browsing in general. The second approach for user identification is using cookies. Cookies are the data sent to the client by the server. The data is then locally stored in cookies, which are sent back to the server with each request. However, there are two problems with this method: first, the users can lock the use of cookies, and the server after that can't save information locally on the user's device. After that, the cookies could be deleted by the user, which means that this technique is not always trustworthy. The proposed approach for identifying individual users in this paper depends on observing their IP addresses. There is a new user if there is a new IP. If the IP address remains the same but there is a difference in the operating system or browsing software, the logical presumption is that every different agent type for an IP address appears as a different user. If all the IP addresses, browsers, and operating systems are the same, the referrer information must be taken into account. And if therefore URI field is empty, a new user session is detected if the URL in the refer URI field has not previously been accessed, or if there is a long gap in time (often more than 10 min) between the accessing time of this record and the preceding one [24].

The goal of session identification is to partition each user's page access into individual sessions. Therefore, a session is a collection of web pages that the same user accesses within a single website visit [25]. The user may have one or more sessions at a given time. The sessions are utilized as data vectors in many tasks like classification, prediction, clustering into sets, etc. Generally, there are two methods that can be utilized for determining the user session. These methods are time-oriented or structure-oriented [26]. Time-oriented builds on the timestamps, or date and time of the request, in the server log file. In the time-oriented session, there are two points for judging the time-oriented session: The first is the difference between the first request and the last request, which must be less than or equal to 30 min. The second is the difference between the first request and the next request, which must be less than or equal to 10 minutes. The structure-oriented

session capture is in the referrer fields of the server logs. The structure-oriented session is captured in the referrer fields of the server logs. While the structure-oriented method is based on referrer fields that are presently open or that the user presently logins too, this means that it belongs to one "open" constructed session [27]. The proposed model suggests the rules below to identify the user session.

- A new session is created whenever a new user logs in.
- A new session is created if the referrer URL field has never been accessed or is empty during a single-user session.
- It is presumed that the user is beginning a new session if the interval between page requests overtakes a specific limit (30 min).

Pattern discovery is the primary task of web mining. Once users and sessions are identified, various data mining techniques, for example statistical analysis [14], association rules [15], sequential patterns [16], clustering [17], classification [18], and dependency modeling [19], are utilized for discovering various groups of matching patterns. Statistical techniques can be applied to gain knowledge about users of a website through the analysis of the session file. A variety of descriptive statistical analyses (frequency, mean, median, etc.) can be performed on variables such as page views, viewing time, and navigation path length. Association rule generation is utilized to associate pages that are frequently referenced with each other in a single server session. For instance, the user searching for journals in any domain might also be interested in conferences, seminars, and workshops in the same domain. The aim of the sequential pattern discovery technique is to discover inter-session patterns in which a group of items is followed by other items in a time-ordered set of sessions. The main goal of clustering analysis in web usage mining is to discover the cluster of users, pages, or sessions, wherever every cluster represents a set of objects with common interests or characteristics. Clustering of users creates a group of users with the same browsing patterns, while clustering of pages highlights and clubs' pages with common or similar content. Classification is the task of representing a data element in one of the classes that have previously been obtained. The data is segregated by comparing the best-qualified features of a class. Dependency modeling is used to create a model that can represent important dependencies between the different variables in a web field. Among these techniques, clustering is the most frequently used. The most significant usage of clustering in web usage mining is getting sets that have mutual interests and behaviors by analyzing the data gathered on web servers.

The final step in web usage mining is pattern analysis. Sometimes the outputs of knowledge mining algorithms are probably not appropriate for human usage directly, and therefore there is a need to improve techniques and tools that need to be better understood by the analyst. These algorithms include fields like statistics, graphics and visualization, usability analysis, and database querying. Pattern analysis performs two processes: validation and interpretation. The validation process eliminates the irrelevant rules (patterns) and extracts the interesting rules from the output of the pattern discovery process. The mining algorithms' outputs are oftentimes mathematical formulas and not appropriate for explicit human explanations. Therefore, the interpretation process is used to explain this output in another form.

III. The Proposed Model

The proposed model is intended to provide a helpful information extraction technique from web log files and use it to realize clusters from related pages to provide assistance for web users in their web navigation. The proposed model is mainly formed by three phases: the pre-processing phase, the pattern discovery phase, and the pattern analysis phase. The functions of the three phases are interpreted as follows: The pre-processing phase analyses the records in the log file to remove irrelevant and redundant data and identify the users' sessions. The pattern discovery phase has two main stages for extracting a set of accurate and comprehensible rules. The first stage is the session cluster technique, and the second stage is the rule extraction technique. The cluster technique is used to discover sets of users' sessions that share common interests and behaviors. Consequently, each user session can belong to one of the classes (clusters). Thus, the input unsupervised data is transformed into supervised data and used as input for the next stage. The rule extraction technique uses the pages of each user session and the corresponding class as the input and output patterns for training the ANN. If the ANN uses a sigmoid activation function, then every output node of the ANN can be represented by an exponential function that depends on the values of session pages. For extracting a set of rules that belongs to a specific cluster, the general exponential function corresponding to this cluster is created, and the GA is utilized to extract the optimal values of pages that maximize this function. Then, these values of pages are converted into a group of rules that belong to a specific class according to a certain threshold level. The pattern analysis phase finds valuable mining results and gives a detailed interpretation according to the discovered rules. Finally, the recommended pages for each cluster can be extracted. Fig. 1 shows a concise proposed framework for web usage mining. The three phases of the proposed model can be explained as follows.

A. Pre-processing Phase

A web log file is a log file that a web server automatically creates and maintains. There are various forms of web log files, such as W3C, NASA, and IIS log files. Analyzing these log files provides useful information about the user. The basic aim of data pre-processing is to eliminate noisy and pointless data and decrease the amount of data available for the pattern discovery phase. The following algorithm is used for data cleansing, as shown in Fig. 2.

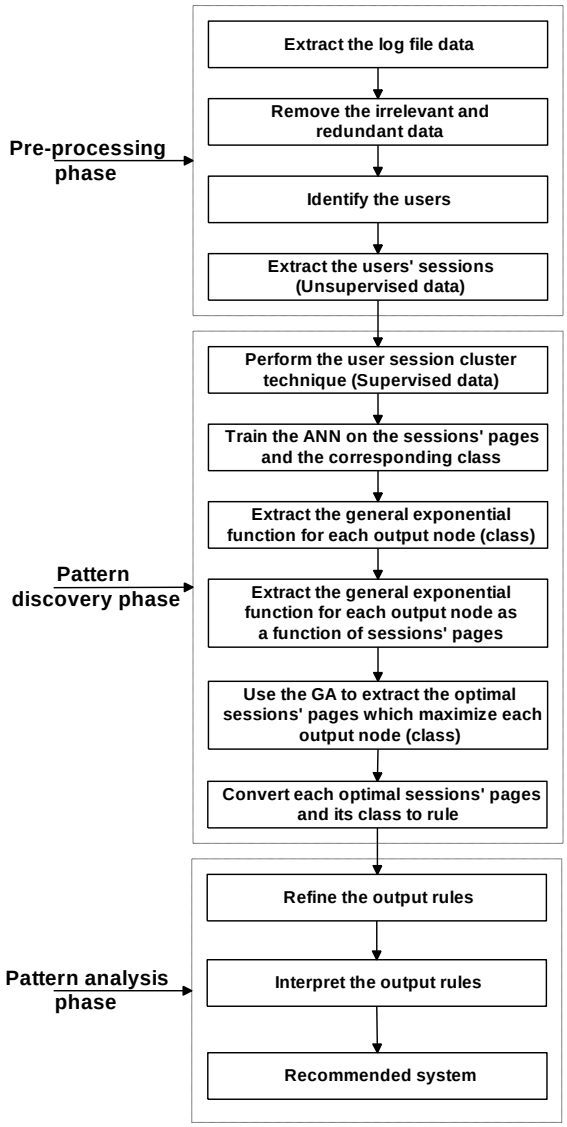

Figure 1. The proposed framework for web usage mining.

Step 1: Read records from the log file.

Step 2: Remove the records which have image files such as .jpg, .gif, .css and all folders which contain images.

Step 3: Remove the records which have HTTP status codes over 299 or fewer than 200.

Step 4: Remove the records which have request method other than GET and POST.

Step 5: Repeat the steps from 2 to 4 until EOF (Web Log File).

Figure 2. Data cleaning algorithm.

After accurate data cleaning, the step of identifying the user is completed. The person who tries to access the web pages from the webserver is called the user. The below algorithm describes the rules for user identification, as shown Fig. 3.

A user session is a collection of pages that a user visits during a single visit to a website. In this paper, the time-oriented method is utilized to determine the user session. The following algorithm suggests the rules used to identify the user session, as shown in Fig. 4.

Step 1: Read records from web log file.

Step 2: User's IP addresses of two consecutive entries are compared.

Step 3: If (IP address is the same) then

Check user's browser and operating system.

If both are different then

Consider new user

Else

Check the referrer URI field.

If the URL in the referrer field hasn't been accessed previously then

consider new user

Else

Consider same user

End if

End if

End if

Step 4: Repeat the steps from 2 to 3 until EOF (web log file).

Figure 3. User identification algorithm.

Step 1: Read records from web log file.

Step 2: If (

There is a new IP user OR

The referrer URL field is not accessed previously or empty OR

The time between page requests > 30 minutes

)

Then it is assumed that the user is starting a new session.

Step 3: Repeat step 2 until EOF (web log file).

Figure 4. Session identification algorithm.

B. Pattern Discovery Phase

The pattern discovery phase is classified into two techniques: The first one is the session cluster technique, and the second is the rule extraction technique. The next sections provide a description of each technique.

1) The session cluster technique

Clustering algorithms are the more dominant approaches that have been utilized for pattern discovery from web data. Clustering is intended to split a data set into different groups whose components are similar to each other. The algorithms for users' session clustering may be classified into two approaches: similarity-based clustering and model-based clustering algorithms. The similarity-based clustering algorithms don't need any assumptions about the data's probability structure. This only needs a similarity function realized on the data pairs [28].

Web session clustering is a technique used in web mining tasks that gathers web sessions based on similarity, with each group having the highest intra-group similarity and the lowest inter-group similarity. This process can be used in website management, web personalization, developing web recommender systems, etc. The clustering accuracy of web sessions is exceptionally dependent on the similarity measure specific to the items to be clustered [29]. Hierarchical and partitional techniques are the most

common indicators that belong to the similarity-based approach [30].

The hierarchical techniques are dependent on measures of distance (similarity) between clusters. Generally, differences among the algorithms that use hierarchical clustering arise mainly because of the several ways of defining the distance between two sessions, such as Euclidean, Cosine, or Manhattan distances [31]. The hierarchical technique merges those two clusters that are the nearest to create a smaller number of clusters. This happens again, each time merging the two closest clusters until just one cluster exists for all the data points. The result is represented by a tree of clusters, which illustrates the relations among them. The data set is decomposed into a predetermined set of disjoint clusters by using a partition-based clustering scheme; therefore, the members in each cluster are identical as much as possible. K-means [32] and k-medoids [33] are examples of methods used for typical partitioning. In the k-means algorithm, every cluster is represented by the mean value of the data points in the cluster, known as the centroid of the cluster. In the k-medoids algorithm, every cluster is represented by one of the data points located near the center of the cluster, called the medoid of the cluster [34].

Model-based clustering algorithms are often utilized to cluster similar users' sessions in order to define web site access behaviors. These algorithms assume a probability model for each cluster and attempt to best fit the data to the assumed model. Where every cluster contains a data-generating model with various parameters for every cluster. The clustering process aimed to retrieve the model parameters and assign sessions to the cluster where the model best described them. The most widely used model-based algorithm is the Expectation-Maximization (EM) algorithm, which is utilized to select associations among users and pages [35] and also for provisioning user profiles [36]. Although similarity-based clustering algorithms are computationally more complex than model-based clustering approaches, they have shown their capacity for producing more effective web session clustering results [37].

TABLE I. THE SESSIONS REPRESENTATION

	P_1	P_2	P_3	P_i	P_n
S_1					
S_2					
S_3					
S_j	P_{1j}	P_{2j}	P_{3j}	P_{ij}	P_{nj}
S_m					

The proposed clustering algorithm's objective is to produce user session data in terms of page views. In other words, the page view vectors are the expression of the user sessions. Given "*n*" web pages in a website and "*m*" web users visiting the website during a period of time, after data preprocessing, we can build up a collection of "*m*" sessions as $S = (S_1, S_2, \ldots, S_j, \ldots, S_m)$. Each session includes "*n*" pages as $P = (P_1, P_2, \ldots, P_i, \ldots, P_n)$. Thus, the web session data could be shown as an "$m \times n$" session-page view binary matrix, where $P_{ij} = 1$ means the page view "P_i" is visited in the user session "S_j", and $P_{ij} = 0$ represents the page which has not been visited in the user session as shown in Table I. These types of data transformations can also be known as page identification and user sectorization processes [38].

The task in this section is to compile the session instances into meaningful session classes. Each session instance is considered a session class, and the distance between them is calculated. Then, the session groups are merged based on their intra-group distances, and group distances are updated correspondingly. The distance between two session instances S_i and S_j can be calculated by comparing each page of S_i with the corresponding page S_j. The comparison results are binary values that present a local similarity measure, $\mathrm{L\_Sim}(P_{ni}, P_{nj})$, according to the follows in Eq. (1):

$$\mathrm{L\_Sim}(P_{ni}, P_{nj}) = \begin{cases} 1, \text{if } P_{ni} = P_{nj} = 1 \\ 0, \text{if } P_{ni} = P_{nj} = 0 \text{ or } P_{ni} \neq P_{nj} \end{cases} \quad (1)$$

The global similarity measure $\mathrm{G\_Sim}(S_i, S_j)$ among all pages of user sessions can be calculated as follows in Eq. (2):

$$\mathrm{G\_Sim}(\mathrm{S_i}, \mathrm{S_j}) = \frac{\sum_{\mathrm{n=1}}^{n} \mathrm{L\_Sim}(\mathrm{P_{ni}}, \mathrm{P_{nj}})}{K} \quad (2)$$

where "*K*" is the maximum number of visited pages in S_i or S_j.

After the global similarity between all sessions is calculated, some of them are illuminated according to the value of the similarity threshold ($\mathrm{Sim_{th}}$). Consequently, the remaining connections between the sessions can be used to set up the required clusters. Once we obtain the clusters, we can identify each user session that belongs to one of the extracted clusters (supervised data). The following illustrative example aims to illustrate how simple the proposed clustering algorithm is. Assume that there are 7 sessions and 10 pages on the website under study. The web session data are shown in Table II.

TABLE II. SESSIONS AND THEIR PAGES

No.	Session Pages
S_1	$\{P_1P_3P_6P_5P_{10}\}$
S_2	$\{P_4P_9P_6P_{10}P_7\}$
S_3	$\{P_7P_6P_5P_4P_{10}P_9\}$
S_4	$\{P_1P_3P_6P_5P_{10}P_9P_2\}$
S_5	$\{P_1P_{10}P_8P_2P_5P_3P_6\}$
S_6	$\{P_9P_{10}P_6P_3P_2P_1\}$
S_7	$\{P_1P_5P_6P_3\}$

The pages of sessions can be represented by a binary form where the visited pages are replaced by "1" and other pages are replaced by "0", as shown in Table III.

TABLE III. SESSIONS' PAGES REPRESENTATION

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
S_1	1	0	1	0	1	1	0	0	0	1
S_2	0	0	0	1	0	1	1	0	1	1
S_3	0	0	0	1	1	1	1	0	1	1
S_4	1	1	1	0	1	1	0	0	1	1
S_5	1	1	1	0	1	1	0	1	0	1
S_6	1	1	1	0	0	1	0	0	1	1
S_7	1	0	1	0	1	1	0	0	0	0

The global similarities between all sessions are calculated according to Eq. (2), and the results are shown in Table IV.

TABLE IV. THE SIMILARITIES MATRIX OF SESSIONS

	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1	1.00	0.40	0.50	0.714	0.714	0.667	0.80
S_2	0.40	1.00	0.833	0.4285	0.285	0.50	0.20
S_3	0.50	0.833	1.00	0.571	0.428	0.50	0.333
S_4	0.714	0.4285	0.571	1.00	0.857	0.857	0.571
S_5	0.714	0.285	0.428	0.857	1.00	0.714	0.571
S_6	0.667	0.50	0.50	0.857	0.714	1.00	0.50
S_7	0.80	0.20	0.333	0.571	0.571	0.50	1.00

Assume that the similarity threshold is $\mathrm{Sim}_{th} = 0.8$, and then the similarities between the sessions that are less than or equal to 0.8 are neglected. Therefore, the similarities matrix of the sessions is modified as shown in Table V.

TABLE V. THE MODIFIED SIMILARITIES MATRIX ($\mathrm{Sim}_{th} = 0.8$)

	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1	1.00						0.80
S_2		1.00	0.833				
S_3		0.833	1.00				
S_4				1.00	0.857	0.857	
S_5				0.857	1.00		
S_6				0.857		1.00	
S_7	0.80						1.00

The proposed clusters can be built by taking each session and determining the other sessions that are connected to it. For example, S_1 is connected to S_7 only, S_2 is connected to S_3 only, and S_4 is connected to S_5 and S_6. Consequently, the results indicate that there are three clusters as follows:

Cluster 1 = { S_1, S_7 }

Cluster 2 = { S_2, S_3 }

Cluster 3 = { S_4, S_5, S_6 }

Otherwise, the similarity matrix in Table IV can be used to illustrate the clustering of sessions in a graphical model.

In graph-based approaches, data is represented as a graph, with nodes representing users and edges representing interactions (similarities) between users [8, 21, 22]. The graphical representation of the proposed clustering technique is shown in Figs. 5–7:

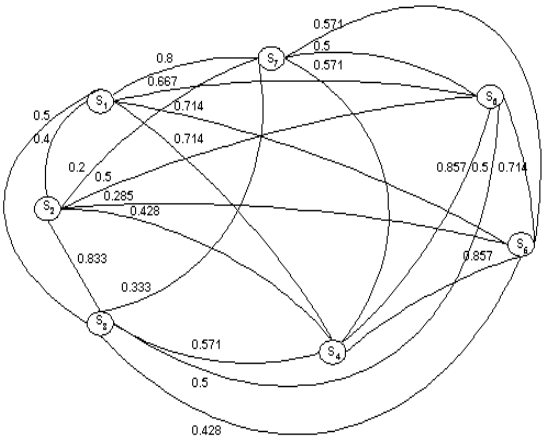

Figure 5. The relationship between sessions.

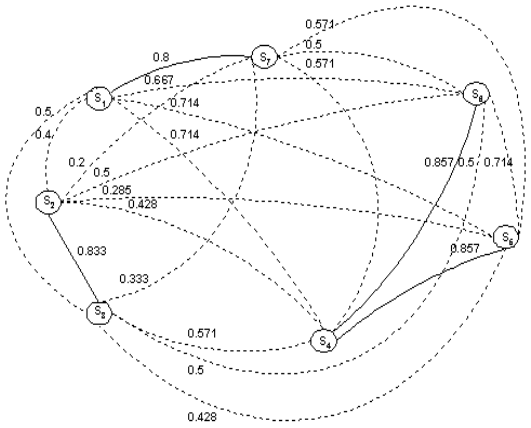

Figure 6. The connections removed according to the Sim_{th}.

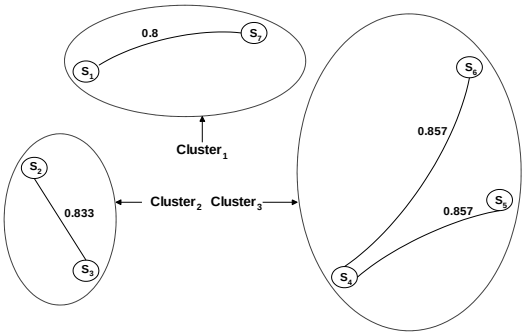

Figure 7. The final results of clusters.

2) The rule extraction technique

The association rules find out the associations and correlations between elements where the existence of an element or a set of elements in a transaction means the

presence of other elements. One of the most common implementations of association rules in web usage mining comes from the relationship between user visits and particular website navigational patterns. Therefore, the present paper introduces a new technique for extracting a set of precise and understandable rules from the clustering users' sessions. The proposed technique extracts the rules from a trained Artificial Neural Network (ANN) utilizing the Genetic Algorithm (GA). The input and output vectors used to train the ANN are the pages of each user session and the corresponding class of the session. Each ANN input unit usually corresponds to a single user session page, and each output unit corresponds to a session class. Each input vector is represented by binary elements equal to the number of pages on the website. Every item of the input vector equals one if its corresponding page is visited; otherwise, it will be represented by zero. Each output vector is represented by a number of binary elements equal to the number of different classes. The element which belongs to a specific class equals 1, while all the other elements in the vector equal 0. ANN is trained with a view to achieving a low error level. Once the ANN is trained, the weights between the input-hidden layer (W_{ij}) and hidden-output layer (W_{jk}) can be extracted. So, each output node of an ANN can be represented as a function of the input pages' values and the extracted weights between the layers. If the sigmoid function is used as the activation function in hidden and output nodes, then the output function of each output node can be represented as follows in Eq. (3):

$$\psi_k = \left\{ 1/\left(1+e^{-\left[\sum_{j=1}^{M} W_{jk} \cdot \left(\frac{1}{1+e^{-[\sum_{i=1}^{N} P_i . W_{ij}]}} \right) \right]} \right) \right\} \quad (3)$$

where:

P_i : The i^{th} page of the web site, (i = 1, ..., N).

N : The number of web site pages.

W_{ij} : The weights between the i^{th} input node and the j^{th} hidden node, (j = 1, …, M).

M : The number of hidden nodes.

W_{jk} : The weights between the j^{th} hidden node and the k^{th} output node, (K= 1, …, O).

O : The number of output nodes.

ψ_k : The output value of the k^{th} output node.

Each output function ψ_k is a non-linear exponential function and its maximum output value equals 1. Depending on this, for getting a relation (rule) between the input pages relating to a specific cluster (class), one must find the input values of pages (P_i) which maximizes ψ_k. Thus, this is an optimization problem and the GA can be utilized for getting the optimal values (chromosome) of the input pages of the user session which maximize the output function for each class.

C. Pattern Analysis Phase

The extracted chromosome must be decoded to find the corresponding rule as following; i) The extracted chromosome is split into N elements (bits), ii) Each element represents one page of the tested web site, iii) The page is visited if the corresponding bit in the optimal chromosome equals one and vice versa, iv) The operator "AND" are utilized to correlate the visited pages, v) The extracted rules must be refined according to two levels. The first level removes redundant rules. For instance, given two rules R_1 and R_2, if $R_1 : T_1 \Rightarrow C$ and $R_2 : T_1 \wedge T_2 \Rightarrow C$, we can say that the second rule R_2 is redundant. The second level keeps only the rules that have fitness values higher than the fitness threshold level (F_{thr}). For example, given a rule R_1, if $R_1 : T_1 \Rightarrow C$ and has a fitness value F_1, then the rule R_1 is generated if $F_1 \geq F_{thr}$.

The following example illustrates how the proposed rule extraction technique is used to extract potentially useful knowledge from the clustering sessions.

As a result of applying the cluster technique, we can identify each user session that belongs to one of the extracted clusters, as shown in Table VI.

TABLE VI. THE USERS' SESSIONS AND THE CORRESPONDING CLASS

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	C_1	C_2	C_3
S_1	1	0	1	0	1	1	0	0	0	1	1	0	0
S_2	0	0	0	1	0	1	1	0	1	1	0	1	0
S_3	0	0	0	1	1	1	1	0	1	1	0	1	0
S_4	1	1	1	0	1	1	0	0	1	1	0	0	1
S_5	1	1	1	0	1	1	0	1	0	1	0	0	1
S_6	1	1	1	0	0	1	0	0	1	1	0	0	1
S_7	1	0	1	0	1	1	0	0	0	0	1	0	0

The ANN is trained on the 7 input user sessions, $f(P_I)$, and the corresponding output classes, $f(C_I)$. The parameters of the ANN are adjusted as follows: The number of input nodes = 10, the number of output nodes = 3, the number of hidden nodes = 5, the learning rate coefficient= 0.286, the momentum coefficient= 0.764, the number of iterations = 30,000, and the allowable error = 0.0000001. Once the ANN is trained, the weights between the layers are extracted and the function ψ_k is generated. Therefore, the GA is utilized to extract the optimal values (chromosome) of the input Ps of the user session which maximize the output function for each class (ψ_k). The parameters of the GA are adjusted as follows: The population size = 10, the number of generations 1350, the

crossover rate = 0.31, and the mutation rate = 0.004. The output chromosomes for each target class are sorted from up to down according to their fitness values until the threshold level = 0.9996. The extracted rules are refined to remove the redundant rules and keep the rules that have fitness values higher than the fitness threshold level. The final set of accurate and comprehensive rules is shown as follows:

If P_1 =1 and P_3 =1 and P_5 =1 and P_6 = 1: Then $Class_1$ Fitness = 0.99998

If P_4 =1: Then $Class_2$ Fitness = 0.99996

If P_7 =1: Then $Class_2$ Fitness = 0.99995

If P_2 =1: Then $Class_3$ Fitness = 0.99997

If P_8 =1: Then $Class_3$ Fitness = 0.99991

If P_1 =1 and P_2 =1: Then $Class_3$ Fitness = 0.99983

If P_1 =1 and P_9 =1: Then $Class_3$ Fitness = 0.99974

If P_3 =1 and P_9 =1: Then Fitness = 0.99967

IV. Recommendation System

The main objective of a web recommender system is to automatically filter and sort information about products or services for a user in a broad web information repository depending on the relationship between the user's needs and the products or services, with no manual effort from the user [39]. The web recommendation system is one of the web personalization techniques, and it is divided into two groups: Content-Based (CB) recommendation [40–43] and Collaborative Filtering (CF) recommendation system [44–47]. Content-based recommended systems are a classifying method derivative of machine learning research. This method analyses item specifications to identify items that match the user profile. The user's profile is dependent on elements they've previously liked or on specific interests they've selected. The CB recommender systems use supervised machine learning in order to create a classifier to differentiate between items likely to be of interest to the user and those that are likely to be uninteresting. The CB recommendation systems may be utilized in various domains, such as recommending websites, news stories, hotels, TV shows, and products for sale. However, CB recommender systems have several shortcomings. One of the shortcomings of content-based recommendation systems is that they can't execute in fields where there is no content correlating with elements or where the content is hard to analyze. Whereas the CF recommendation system is the more effective recommendation technique and is utilized in a variety of various applications [48]. It depends on the notion that a person who has made previous choices is likely to make the same choices in the future. The CF recommendation system is utilized to create a recommendation for a user by finding a group of users named "neighborhood" that have tastes similar to those of the target user. The similarity of elements is identified by the similarity of the classifications of certain elements by the users who have classified both elements. The present work provides a proposed model for developing a collaborative filtering recommendation system using a rule mining technique. The rule mining technique extracts precise and understandable rules from the trained neural networks using a genetic algorithm. Thus, the extracted rules from the preceding section can be utilized to build the proposed web recommendation system. For each group of rules that belongs to a specific cluster, the dominant pages in all rules are used as recommendation pages for this cluster. Therefore, the recommended web pages for the three clusters in the previous example are shown in Table VII.

TABLE VII. The Recommended Pages for the Three Clusters

Cluster No.	Recommended Pages
$Cluster_1$	$\{P_1P_3P_5P_6\}$
$Cluster_2$	$\{P_4P_7\}$
$Cluster_3$	$\{P_1P_2P_3P_8P_9\}$

For evaluating the efficacy of the proposed framework, three criteria are used: recall, precision, and f-measure [49]. The recall is the part of all pertinent elements that are recommended and represents the reliability of the proposed recommendation system, while the precision is the part of all recommended elements that are relevant. Therefore, each item in the recommendation set can be either relevant or irrelevant to the user, and the confusion matrix can be written as shown in Table VIII.

TABLE VIII. The Confusion Matrix for The Recommended System

	Recommended	Not Recommended
Relevant	A	C
Not Relevant	B	D

$$\Pr ecision = \frac{A}{A+B} \tag{4}$$

$$\mathrm{Re}call = \frac{A}{A+C} \tag{5}$$

Due to the nature of these two-classification metrics, there is a tradeoff between recall and precision; as the number of recommended items increases, the recall tends to increase while the precision decreases. Then, a commonly utilized combination metric called the F1 metric, which gives equal weight to both recall and precision, was utilized for our evaluation. The F1 metric is calculated using the following equation:

$$F-measure = \frac{2\times \mathrm{Re}call \times \Pr ecision}{\mathrm{Re}call + \Pr ecision} \tag{6}$$

V. Experimental Results

For validating the efficiency and effectiveness of the proposed framework, we applied an experiment using the webserver log of the "Effectiveness of a blended learning-based program for developing some English language skills for engineering faculty students" web site at Mansoura University (http://osp.mans.edu.eg/blended-learning/). This web site was designed as an e-learning program for 2nd-year students of the computer department, faculty of engineering, Delta University, Mansoura, Egypt.

The main data source in this experiment is from May 1, 2022, to June 18, 2022. After data cleaning, we had about 6.2 MB of 21,426 cleaned records. After session extraction, we had 2,108 user sessions and 65 distinct pages. These experiments were applied to a 3.0 GHz Pentium CPU with 1024 MB of main memory. The session cluster technique is performed at the similarity threshold level $\mathrm{Sim}_{\mathrm{th}} = 0.73$ for clustering the user sessions. Thus, the user sessions are classified into 37 clusters. Once we obtain the clusters, we can identify each user session that belongs to one of the extracted clusters. Now, the rule extraction technique can be applied to extracting a set of accurate and comprehensible rules from the clustering users' sessions. The final parameters' values of ANN are depicted as follows: there are 65 input nodes, 8 hidden nodes, 37 output nodes, 0.229 learning rate, and 0.648 momentum., the allowable error is set to 0.000001 and the number of iterations is set to 60,000. Once the ANN is trained, the weights between the layers are extracted and the function ψ_k is generated. Therefore, the GA is utilized to extract the optimal values of the input pages of the user session which maximizes the output function for each class (ψ_k). The parameters of the GA are adjusted as follows: population size = 20, count of generations = 35,000, crossover rate = 0.281, and mutation rate = 0.027. The output chromosomes for each target class are ordered from up to down according to their fitness values until the threshold level = 0.99994. The extracted rules are refined to remove redundant rules and keep the rules that have fitness values higher than the fitness threshold level. The web recommendation pages for each cluster are built according to the extracted rules. Fig. 8 shows the precision-recall curve to validate the effectiveness of the proposed recommendation system.

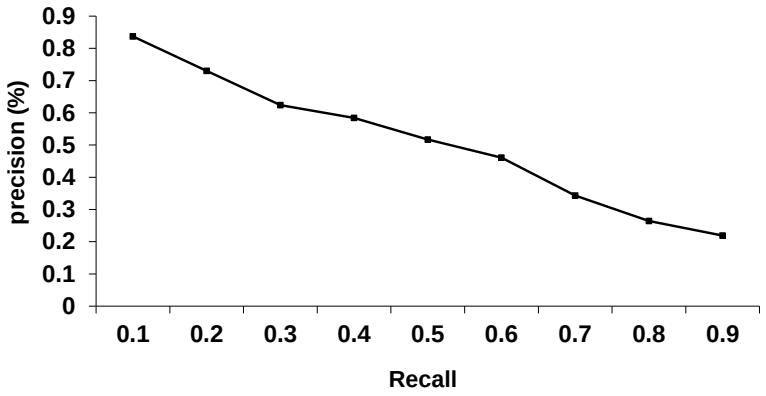

Figure 8. The precision-recall for the proposed recommendation system.

VI. Conclusion

This work introduced a new model for improving a collaborative filtering web recommendation system. It begins with preprocessing and cleaning the used data, then clustering the sessions based on global similarity. The proposed rule extraction technique is based on the ANN and GA and is refined to remove non-interesting rules. The results of the implementation revealed that the proposed web recommendation model performed well in terms of precision and recall measures.

The proposed model introduced a new simple technique for extracting precise and understandable knowledge and achieved higher accuracy than other related systems.

Conflict of Interest

The author declares no conflict of interest.

References

[1] A. Kour *et al*., "Web mining in soft computing relevance and future directions," *International Journal of Electronics Communication and Computer Engineering*, vol. 4, no. 1, pp. 2278–4209, 2013.

[2] B. Harika and T. Sudha, "Identification of user behaviour by web usage mining," *Mathematical Statistician and Engineering Applications*, vol. 71, no. 4, pp. 678–692, 2022.

[3] A. Mebrahtu and B. Srinivasulu, "Web content mining techniques and tools," *International Journal of Computer Science and Mobile Computing*, vol. 6, no. 4, pp. 49–55, April 2017.

[4] D. Bhadoria and P. Sharma, "Review paper on web structure mining," *International Journal of Scientific Engineering and Research (IJSER)*, vol. 4, no. 7, pp. 54–61, July 2016.

[5] S. M. Patil, T. V. Kumar, and H. S. Guruprasad, "Comparative analysis of web usage data using SOM and K-means algorithms," *International Journal of Advanced Research*, vol. 3, no. 10, pp. 486–493, 2015.

[6] R. Shah and S. Jain, "Web mining using cloud computing technology," *International Journal of Scientific Research in Computer Science and Engineering*, vol. 3, no. 2, pp. 21–25, 2015.

[7] V. Sathiyamoorthi and M. Bhaskaran, "Data preprocessing techniques for pre-fetching and caching of web data through proxy server," *International Journal of Computer Science and Network Security*, vol. 11, no. 11, pp. 92–98, November 2011.

[8] A. Prasanth, "Web personalization using web usage mining techniques," *International Journal of Current Engineering and Scientific Research (IJCESR)*, vol. 3, no. 3, pp. 45–49, 2016.

[9] D. Desai, "Website personalization: strategy for user experience design and Development," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 12, pp. 3516–3523, 2021.

[10] D. Patel, A. Patel, and K. Parikh, "Preprocessing algorithm of prediction model for web caching and perfecting," *International Journal of Information Technology and Knowledge Management*, vol. 4, no. 2, pp. 343–345, 2011.

[11] N. Kandpal, R. R. Sinha, and M. S. Shekhawat, "A survey on web usage mining: Process, application and tools," *An International Bi-Annual Journal*, vol. 3, no. 1, pp. 19–25, 2017.

[12] A. Sivakumar and R. Gunasundari, "A survey on data preprocessing techniques for bioinformatics and web usage mining,"

International Journal of Pure and Applied Mathematics, vol. 117, no. 20, pp. 785–793, 2017.
[13] A. A. R. A. Azmi, "Data, text, and web mining for business intelligence: A survey," *International Journal of Data Mining & Knowledge Management Process*, vol. 3, no. 2, pp. 1–21, March 2013.
[14] M. Aldekhail, "Application and significance of web Usage mining in the 21st century: A literature review," *International Journal of Computer Theory and Engineering*, vol. 8, no. 1, pp. 41–47, February 2016.
[15] P. Sharma, D. Yadav, and R. N. Thakur, "Web page ranking using web mining techniques: A comprehensive survey," *Hindawi Mobile Information Systems*, vol. 12, pp. 1–19, May 2022.
[16] S. Sahu, R. Gupta, and A. Dutta, "An analysis of web user behavior using hybrid algorithm based on sequential pattern mining," *International Journal of Applied Engineering Research*, vol. 14, pp. 2339–2346, 2019.
[17] L. Lu and Y. X. Tu, "An improved hierarchical clustering algorithm for performance testing based on user sessions," *Journal of Computers*, vol. 30, no. 5, pp. 145–158, 2019.
[18] H. E. Unal, S. A. Ozel, and I. Unal, "Performance of using tag-based feature sets in web page classification," *Journal of Natural and Applied Sciences*, vol. 22, no. 2, pp. 583–594, 2018.
[19] H. Amirat *et al*., "MyRoute: A graph-dependency based model for real-time route prediction," *Journal of Communications,* vol. 12, no. 12, pp. 668–676, December 2017.
[20] R. Vasudevan, "Neural networks and web mining," *SSRG International Journal of Electronics and Communication Engineering*, vol. 1, no. 1, pp. 9–14, Feb. 2014.
[21] B. Desai, R. Prajapati, and S. Khanna, "Web usage mining: A survey on extracting knowledge through web logs," *International Journal of Scientific Development and Research*, vol. 1, no. 5, pp. 468–472, May 2016.
[22] K. B. Patel and A. R. Patel, "Process of web usage mining to find interesting patterns from web usage data," *International Journal of Computers and Technology*, vol. 3, no. 1, pp. 144–148, August 2012.
[23] B. Chen, W, M. Peng, and J. H. Song, "Sequential pattern mining with multidimensional interval items," *Technical Gazette*, vol. 29, pp. 1220–1229, 2022.
[24] W. A. G. A. Hussain, "Identifying of user behavior from server log file," *Iraqi Journal of Science*, vol. 58, no.2C, pp. 1136–1148, 2017.
[25] S. Kalaivania and K. Shyamalab, "Clustering of web users behavior based on the session identification through web server log file," *International Journal of Control Theory and Applications*, vol. 10, pp. 7–16, 2017.
[26] J. Kapusta *et al*., "User identification in the process of web usage data preprocessing," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 9, pp. 21–32, 2019.
[27] W. Chandrama, P. R. Devale, and R. Murumkar, "Data preprocessing method of web usage mining for data cleaning and identifying user navigational pattern," *International Journal of Innovative Science, Engineering and Technology*, vol. 1, no. 10, pp. 73–77, December 2014.
[28] X. F. He *et al*., "Laplacian regularized gaussian mixture model for data clustering," *Journal of Latex Class Files*, vol. 6, no. 1, pp. 1–14, January 2007.
[29] V. Dogne, A. Jain, and S. Jain, "Evolving trends and its application in web usage mining: A survey," *International Journal of Soft Computing and Engineering*, vol. 4, no. 6, pp. 98–101, January 2015.
[30] N. Huidrom and N. Bagoria, "Clustering techniques for the identification of web user session," *International Journal of Scientific and Research Publications*, vol. 3, no. 1, pp. 1–8, January 2013.
[31] M. H. A. Elhebir and A. Abraham, "Discovering web server logs patterns using clustering and association rules mining," *Journal of Network and Innovative Computing*, vol. 3, pp. 159–167, 2015.
[32] K. Kim and H. Ahn, "A recommender system using ga k-means clustering in an online shopping market," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1200–1209, February 2008.
[33] S. Lomate, "Web personalization recommendation system based on clustering and association rule," *International Journal for Technological Research in Engineering*, vol. 4, no. 2, pp. 263–266, 2016.
[34] D. S. Sisodia, S. Verma, and O. P. Vyas, "A subtractive relational fuzzy c-medoids clustering approach to cluster web user sessions from web server logs," *International Journal of Applied Engineering Research*, vol 12, no. 7, pp. 1142–1150, 2017.
[35] M. H. A. Elhiber and A. Abraham, "Access patterns in web log data: A review," *Journal of Network and Innovative Computing*, vol. 1, pp. 348–355, 2013.
[36] P. Priya and M. Hemalatha, "Mathematical user profiling algorithms for web recommendation based on PLSA and LDA model," *Global Journal of Pure and Applied Mathematics*, vol. 13, no. 9, pp. 6713–6721, 2017.
[37] J. Domenech *et al*., "A comparison of prediction algorithms for prefetching in the current web," *Journal of Web Engineering*, vol. 11, no. 1, pp. 64–78, 2012.
[38] V. G. Manjula and Y. J. Singh, "Web mining for social network analysis: A review, direction and future vision," *ADBU-Journal of Engineering Technology*, vol. 4, no. 1, pp. 14–22, 2016.
[39] R. Shanthi and S. P. Rajagopalan, "A personalized hybrid recommendation procedure for internet shopping support," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 12, pp. 363–372, 2018.
[40] S. Yanga *et al*., "Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive statistical relational learning approach," *Knowledge-Based Systems*, vol. 136, pp. 37–45, 2017.
[41] A. Prasanth, "Intelligent recommendation system using semantic information for web information retrieval," *Advances in Computational Sciences and Technology*, vol 10, pp. 2367–2380, 2017.
[42] D. S. Li *et al*., "Interest-based real-time content recommendation in online social communities," *Knowledge-Based Systems*, vol. 28, pp. 1–12, April 2012.
[43] S. G. Esparza, M. P. O. Mahony and B. Smyth, "Mining the real-time web: A novel approach to product recommendation," *Knowledge-Based Systems*, vol. 29, pp. 3–11, 2012.
[44] B. Alhijawi and G. A. Naymat, "Novel positive multi-layer graph based method for collaborative filtering recommender systems," *Journal of Computer Science and Technology*, vol. 37, pp. 975–990, 2022.
[45] C. Cechinel *et al.*, "Evaluating collaborative filtering recommendations inside large learning object repositories," *Information Processing and Management*, vol. 49, no. 1, pp. 34–50, January 2013.
[46] C. F. Tsai and C. Hung, "Cluster ensembles in collaborative filtering recommendation," *Applied Soft Computing*, vol. 12, no. 4, pp. 1417–1425, April 2012.
[47] S. O. Birim and A. Tumturk, "A novel algorithmic similarity measure for collaborative filtering: A recommendation system based on rating distances," *Academic Platform Journal of Engineering and Smart Systems*, vol. 10, pp. 57–69, 2022.
[48] B. Satish and P. Sunil, "Study and evaluation of user's behavior in e-commerce using data mining," *Research Journal of Recent Sciences*, vol. 1, pp. 375–387, 2012.
[49] M. Kuanr and P. Mohapatra, "Assessment methods for evaluation of recommender systems: A survey," *Foundations of Computing and Decision Sciences*, vol. 46, no. 4, pp. 394–421, 2021.