A NEW APPROACH FOR WEB USAGE MINING USING CASE BASED REASONING

Shiva Asadianfam [a,*], **Hoshang Kolivand** [b], **Sima Asadianfam** [c]

[a] Department of Computer Engineering, Qom Branch, Islamic Azad University, Qom, Iran

[b] Department Computer Science, Liverpool John Moores University, Liverpool, UK

[c] Department of Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran

E-mail: [a]sh_asadianfam@yahoo.com, [b] h.kolivand@ljmu.ac.uk, [c] Sima67_asadianfam@yahoo.com

ABSTRACT

Information overload is a major problem in the current World Wide Web. To tackle this problem, web personalization systems which have the capability to adapt the next set of visited pages to individual users according to their interests and navigational behaviors have been proposed. Personalization processes based on web usage mining include the following three phases: data preparation, pattern discovery, and recommendation. In this study, we presented a new approach for Web Usage Mining using Case Based Reasoning. Case-Based Reasoning techniques are a knowledge-based problem-solving approach which is based on the reuse of previous work experience. Thus, the past experience can be deemed as an efficient guide for solving new problems. The proposed architecture consists of a number of components, namely, basic log preprocessing, pattern discovery methods (By CBR and peer to peer similarity - Clustering - association rules mining methods), and recommendations. One of the issues considered in this study is that there are no recommendations to those who are different from the existing users in the log file. Also, it is one of the challenges facing the recommendations systems. To deal with this problem, Apriori algorithm was designed individually in order to be utilized in presenting recommendations; in other words, in cases where recommendations may be inadequate, using association rules can enhance the overall system performance recommendations. A new method used in this study is clustering algorithms for Nominal web data. Our evaluations show that the proposed method along with Standard case-classified Log provides more effective recommendations for the users than the Logs with no case classification.

KEYWORDS

The next-page recommendation for the users' navigation, Case Based Reasoning,web usage mining Techniques, Web Personalization

1. INTRODUCTION

We live in an era in which human beings generated and pass on the data and information more quickly than any other time in the past. Indeed, nowadays, there is far too information to be analyzed. Thus, the user's resource selection from this large amount of information becomes more and more difficult. Web has very different users and each user may only be interested in a small part of the web. Therefore, users have many problems in finding desired information. Rich sources of information on the web that can give rise to web mining help us find and extract information sources. Web mining is the application of data mining techniques to discover patterns from the world-wide network [1]. Web usage mining (WUM) is the use of the data and techniques associated with Web navigation to discover the patterns used in the Web data, for the purpose of better understanding and applying the requirements of Web-based application programs [2]. The goal of web usage mining is to identify patterns from browsing data, which are seen as user interaction with web sites that reflect user interests [3-5] in order to adapt web sites to user interests [6], create recommender systems [7], personalize information [8, 9], etc. Many research in several areas, such as web usage mining has been undertaken in order to detect the

ways to use the user behavior to create a model of his/her interests implicitly. In the field of Web personalization and recommendation of user navigational pages, user models can only be designed based on the web usage data. It provides a shallow understanding of the patterns, and also the page's content can be used for providing better recommendations. In this section, initially, the research done in the field of web usage mining and other utilized techniques used in recommendation systems are briefly introduced. Then, the studies are scrutinized based on the Case-Based Reasoning. Cristobal Romero et al [10] are proposed a developed architecture for a private system to facilitate the Web mining. The proposed architecture in this paper includes both basic and advanced modes. For making recommendations, in the basic mode, only one sequential mining algorithm is used, and in the advanced mode, multiple sequential mining algorithms are used. V.Sujatha and Punithavalli [11] are used the classification algorithm of LCS (finding the longest sequence) to predict the user's future requests. In the paper [12], a new prediction model based on the hierarchical characteristics of the web site is provided. In the paper [13], a new data mining method named Temporal N-Gram (TNGram) is offered to design prediction models of web user navigation regarding the characteristics of the time evolution (considering the time of exploring) of using the website. And there is a wide range of application for the case-based reasoning method, for instance, as mentioned in paper [14], its application in identifying faults in the systems, as seen in the paper [15], in diagnosing the disease, as pointed out in the paper [16], in scheduling, as discussed in the article [17], in making the marketing plans and so on. In addition, in the paper [18], an acceptable number of researches used this method to solve problems in various fields and demonstrated satisfactory results.

The main approach of this paper is to improve the performance of prediction methods by applying Case-based Reasoning (CBR) method and considering web usage mining based on users' past experiences for predicting the user navigation page. The model proposed in this paper uses algorithms like Peer to Peer similarity - clustering - association rules mining methods, so that it can extract better features from the users and follow them and use the Apriori algorithm to achieve better results in the pattern discovery phase. The necessity of this research is that most users have challenges with choosing the web page from among the massively available web pages. The solution is to provide the recommendations to the user on the choice of the next page for navigation, this is done by examining the similarity of former users and how they navigate to the target user. Therefore, the purpose of this study is to improve the user recommendations that can be used to predict the next navigable page of the user. Due to the advantages stated for the CBR method [19], this method is an effective way of recommending the new page to the users for navigation. In this study, each case in the CBR is comprised of sessions and the approach of various users to scrolling, and according to the general procedure for personalization based on WUM, the model consists of three phases. In the data preparation phase, due to its various records, high volume, and sufficient data the Log File is used for training and testing the phases of such a model is used to web usage mining. In the second phase, namely, the pattern discovery phase, Peer to Peer similarity - clustering - association rules mining methods is used and, finally, in Recommendation phase, recommendations for navigations are offered to the current user.

In this paper, in Section 2, the background information required for the better understanding of the method presented in this paper is discussed. In Section 3, first the precise definition of the problem is expressed. The proposed system architecture, its components, and its functions are explained. Then, the proposed architecture is implemented on a set of standard data, and in Section 4, the results of the implementation are presented and then a list of dos and don'ts is provided based on the findings. Finally, in Section 5, the general conclusion and some suggestions for further research are discussed.

2. MATERIALS AND METHODS

In this part, the background information necessary to understand the method presented in this paper is explained. The user behavior analysis can provide relevant insights into the user customization and personalization of web pages. Individual systems must be able to anticipate users' needs based on the user's previous interactions. So, a personal work can be considered as a prediction problem [20]. Also, the reason of using CBR method in the next-page recommendation for the users' navigation is assumed that similar problems maybe have similar solutions. Thus, new problems can be solved using the experience of the previous issues. Our main achievement is to offer some solutions to problems via a Case-Based Reasoning method quickly; In other words, we eliminate the time required to derive the answers, which is regarded as a significant advantage of this method. At first, web usage mining is expounded, and then the Case-Based Reasoning method is explained.

2.1. WEB USAGE MINING

In web usage mining, the records of the reports on the web are reviewed in order to discover user access patterns of web pages. While web content mining and web structure mining use the raw and real Web data, Web usage mining analyses the secondary data derived from the users' interactions with the web. Web usage data include information on the available reports to Web servers, probes reports, user profiles, user information, transactions or sessions, cookies, user queries, lists of the users' desired addresses, mouse clicks, and any other data that can be considered as the result of a user's actions. Analyzing and finding orders in the records of web reports can recognize the potential customers in e-commerce, can enhance the quality of information services for the Internet users, and increase the efficiency of web servers systems [2, 21].

2.2. CASE-BASED REASONING

Case-based Reasoning techniques based on a knowledge-based problem solving approach functions based on the reuse of the previous experiences and have emerged from research in cognitive sciences [3]. Unlike traditional knowledge-based techniques, in order to solve a particular problem, the CBR performs based on the experience gathered in the case base. In solving problems via this approach, the main activities include [22]:

- Retrieve "the similar samples" to the new issue.
- Use the similar issue response retrieved for preparing a proposed answer to the new issue.
- Revisions of the proposed response, if there is a discrepancy in the new and retrieval issue.
- Keep new samples (the new question and its response) for the future use.

It should be noted that the CBR does not offer a definitive solution; however, it provides some assumptions and ideas for the passage of the solution space. The CBR technique is useful when there is no full understanding of the response space, and it is feasible to have the repetitive issues with similar characteristics. In the paper [23], it is pointed out that the success of the case-based reasoning depends on the quality of the knowledge and the accuracy of the reasoning. So, according to the aforementioned characteristics, the issue of the recommendations on the next navigation page of the user can be considered as one of the applications that provide opportunities to exploit the CBR.

3. THE PROPOSED METHOD

In this section, a new proposed method of web usage mining with the Case-Based Reasoning is provided to recommend the user's next navigational pages via the user behavior records in the log file. At first, a precise definition of the problem is presented, and then, a checklist of the desired features of the model is proposed. Afterwards, the system architecture is propounded and its components and the way they work are elaborated.

3.1. THE STATEMENT OF THE PROBLEM

This paper was conducted to introduce a method to predict the next navigational page of users based on the behavior of the previous users. This method has the following characteristics:

- The model is designed via web usage mining techniques.
- The model is created based on the flow behavior of users in a certain period of time, e.g., two days in an EPA log file.
- In the prediction process, the CBR method is used.
- MSNBC is utilized to improve results of the log file.
- In the absence of similar users for recommendations, association rules are applied.

In order to adopt a more formal view, suppose the user U in the period of T has visited the Web pages, such as {p1, p2, p3, p4, p5, p7} and has had the sessions of {s1, s2, ..., sm}. The purpose is the recommendations on the next page scrolling with the CBR method according to the users' behavior and their navigational paths. To put it in other terms, if the user V visited the web pages of {p1, p2, p3, p4} in the period of T1, in the next scrolling, the page {p5} is recommended with respect to the user behavior similar to that user's behavior (means U), so that the proposed navigational pages are properly chosen (i.e., having a good precision and recall). In the absence of the proposal to the user, the most widely used association rules mining algorithm (Apriori) is used.

3.2. THE PROPOSED SYSTEM ARCHITECTURE

The general trend of personalization based on WUM is comprised of three phases: data preparation, pattern discovery, and recommendation. Initially, web log files are converted to the appropriate format, and using data mining techniques, the pattern discovery is applied on them, and finally some good offers are recommended to the users [10]. In this study, the same procedure is used. The proposed system architecture is shown in Figure1. As seen in the Figure1, the proposed system consists of a number of components, namely, basic Data preparation, Pattern discovery methods (By CBR and peer to peer similarity - Clustering - association rules mining methods), and Recommendations. Each phase is described in the following sections.

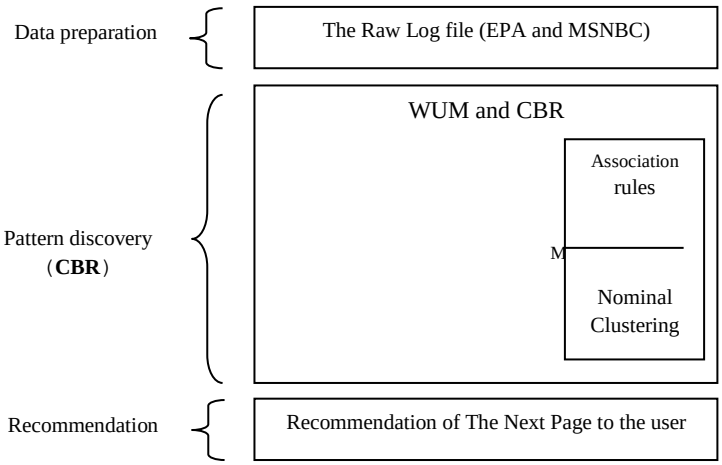


Figure 1- Overview of the proposed framework

3.3. DATA SETS

One of the major problems of research in the domain of the Web personalization is the lack of standard datasets. Due to the issues being private, the web servers' records are not usually available. This problem is more accentuated when the information content is also required. All of the available data are gathered several years ago. In this paper, the log files of EPA and MSNBC are used and the results of each of the two data sets are presented in separate sections. Initially, the EPA log file has been examined,

and then, in order to enhance the results, the MSNBC log file, each page of which is deemed as a part of the proposed seventeen pages classification, is used. To better understand the issue, the process of the proposed architecture based on the web usage mining along with the case-based reasoning method is depicted in Figure 2 in more detail.

3.4. EPA LOG FILE

The EPA Log file is located on the website http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html including 47,748 requests. The Log file size is 4.4MB. Its format is also similar to the typical log files. Each line in this text file shows a specific operation requested by the browser of a user and received by an EPA web server in Research Triangle Park, North Carolina. Each record includes the IP address, Date / Time field, HTTP request, the status code field, and the transfer size field [24].

3.4.1. DATA PREPARATION

Preprocessing due to a defect data in the process of the web usage mining data is very complicated. This phase involves the intensive task of all aspects of the final preparation phase of the data set which are going to be used in all the subsequent stages from the raw data. The selection of cases and variables needed for the analysis, and, if necessary, the modifications on certain variables can be performed in this phase. In other words, the raw web log files do not seem to have an appropriate format for the data mining. Therefore, the data preprocessing is required to be performed [24]. In this study, for the pre-processing operations and the implementation of the process, Matlab R2010A software is used. The most common pre-processing functions include data cleaning and filtering, the user identification, and the session identification. In the following sections, each of the steps will be described. To this end, the text file is put in the form of a matrix, and then the IP address, Date / Time field, HTTP request, the status code field and the transfer size are derived, and so the pre-processing operations begin.

- **Information cleaning and filtering:** In the first step of the pre-processing phase, at first, the suffix of the pages is checked and the irrelevant suffixes are identified. Then, the records including the requests with the specified suffixes such as .gif and .jpeg are removed. In the next stage, the Null requests in the specified log file are detected, and these NULL requests are cleaned from the data set. The status codes other than 200 series, i.e., those requests that fail are deleted. A part of the EPA log file after being cleaned and filtered is shown in Figure 3. After removing graphic and irrelevant files, via string manipulation functions and commands of Matlab programming language, the date/time variables are extracted from Date / Time field and the request method, URL and the protocol version are extracted from the HTTP request field. Then, each request is converted into a number of cones. The reason of doing so is that we make an attempt to convert all values of the fields of a log file into the same type. As a result, it is easier to analyze a homogenous data set.
- **User Identification:** This step should identify unique users. The reasonable assumption for doing so is that each different agent provides an IP address of a different user [12]. In the used log file, each user has a unique IP address, so, each unique IP address represents a user.

- **User Session Identification:** The purpose of identifying sessions is to divide the access of each user to separate sessions. The simplest approach is to use an expiration time (threshold), meaning

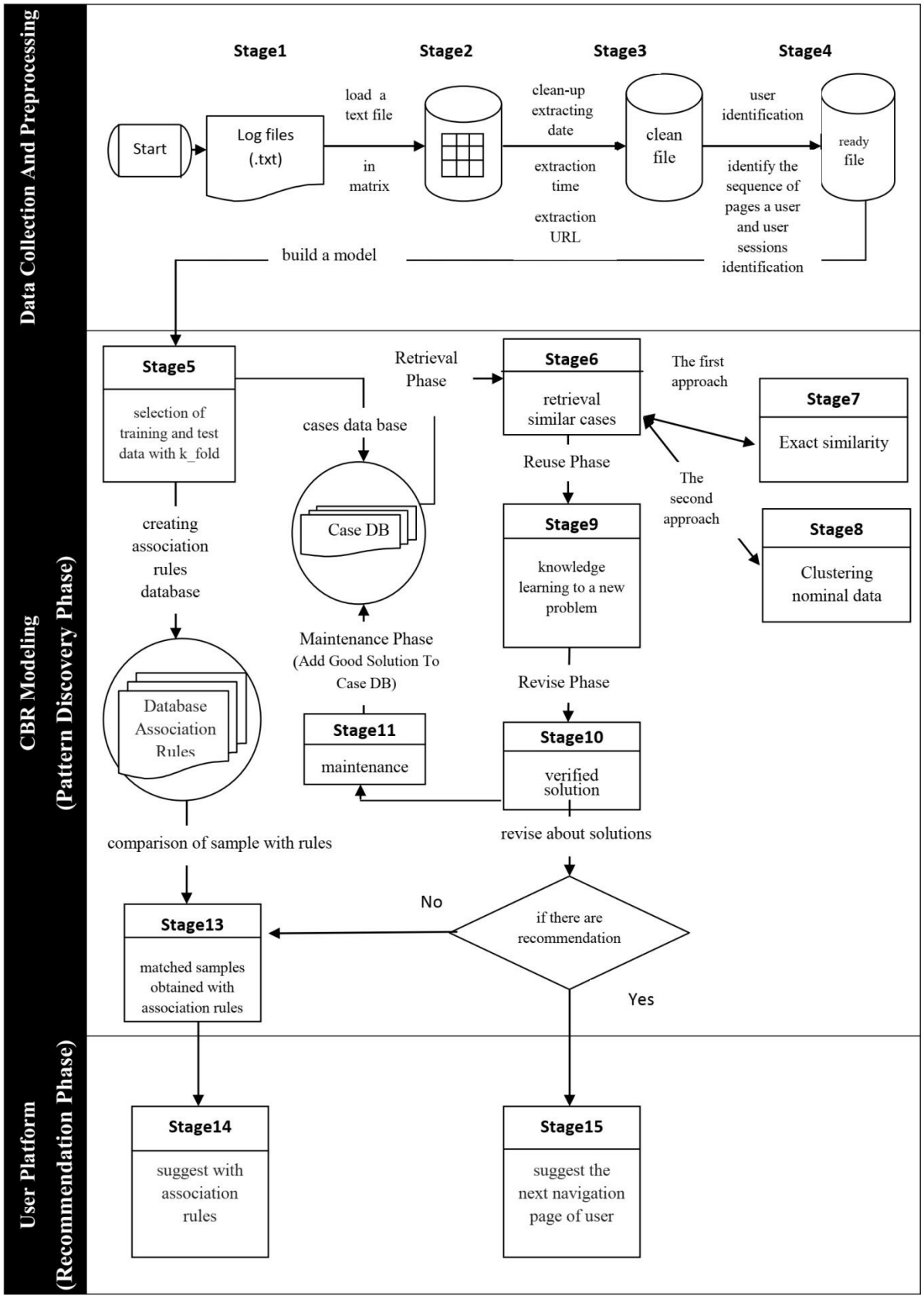
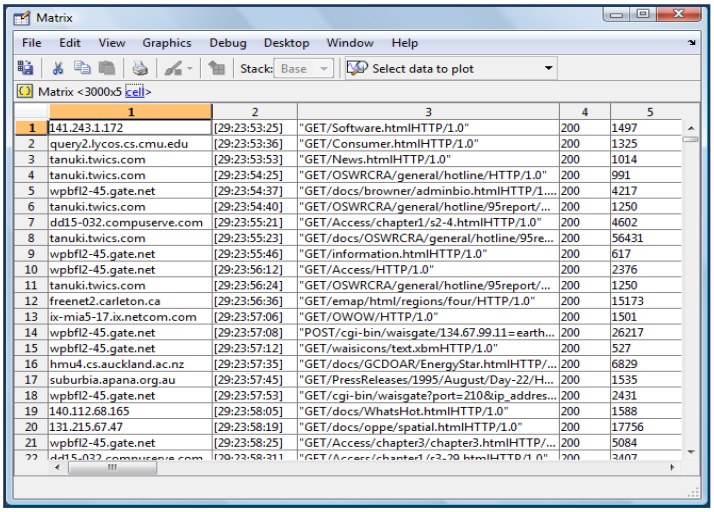


Figure 2 - Process of the proposed architecture based on the web usage mining along with the Case-Based Reasoning in detail

that, if the time spent on a page exceeds a certain threshold, it is assumed that the user has started a new session [25]. So here, the time spent on pages by each user is equal to the difference between the time spent on the next page and the time spent on the previous page by a particular user in seconds. In the

mentioned log file, 30 minutes is used as the completion time of a session and the start time of another session.

Thus, after the preparation phase, the number of records decreased to 9672 records.



	1	2	3	4	5
1	141.243.1.172	[29:23:53:25]	"GET/Software.htmlHTTP/1.0"	200	1497
2	query2.lycos.cs.cmu.edu	[29:23:53:36]	"GET/Consumer.htmlHTTP/1.0"	200	1325
3	tanuki.twics.com	[29:23:53:53]	"GET/News.htmlHTTP/1.0"	200	1014
4	tanuki.twics.com	[29:23:54:25]	"GET/OSWRCRA/general/hotline/HTTP/1.0"	200	991
5	wpbfl2-45.gate.net	[29:23:54:37]	"GET/docs/browner/adminbio.htmlHTTP/1....	200	4217
6	tanuki.twics.com	[29:23:54:40]	"GET/OSWRCRA/general/hotline/95report/...	200	1250
7	dd15-032.compuserve.com	[29:23:55:21]	"GET/Access/chapter1/s2-4.htmlHTTP/1.0"	200	4602
8	tanuki.twics.com	[29:23:55:23]	"GET/docs/OSWRCRA/general/hotline/95re...	200	56431
9	wpbfl2-45.gate.net	[29:23:55:46]	"GET/information.htmlHTTP/1.0"	200	617
10	wpbfl2-45.gate.net	[29:23:56:12]	"GET/Access/HTTP/1.0"	200	2376
11	tanuki.twics.com	[29:23:56:24]	"GET/OSWRCRA/general/hotline/95report/...	200	1250
12	freenet2.carleton.ca	[29:23:56:36]	"GET/emap/html/regions/four/HTTP/1.0"	200	15173
13	ix-mia5-17.ix.netcom.com	[29:23:57:06]	"GET/OWOW/HTTP/1.0"	200	1501
14	wpbfl2-45.gate.net	[29:23:57:08]	"POST/cgi-bin/waisgate/134.67.99.11=earth...	200	26217
15	wpbfl2-45.gate.net	[29:23:57:12]	"GET/waisicons/text.xbmHTTP/1.0"	200	527
16	hmu4.cs.auckland.ac.nz	[29:23:57:35]	"GET/docs/GCDOAR/EnergyStar.htmlHTTP/...	200	6829
17	suburbia.apana.org.au	[29:23:57:45]	"GET/PressReleases/1995/August/Day-22/H...	200	1535
18	wpbfl2-45.gate.net	[29:23:57:53]	"GET/cgi-bin/waisgate?port=210&ip_addres...	200	2431
19	140.112.68.165	[29:23:58:05]	"GET/docs/WhatsHot.htmlHTTP/1.0"	200	1588
20	131.215.67.47	[29:23:58:19]	"GET/docs/oppe/spatial.htmlHTTP/1.0"	200	17756
21	wpbfl2-45.gate.net	[29:23:58:25]	"GET/Access/chapter3/chapter3.htmlHTTP/...	200	5084
22	dd15-032.compuserve.com	[29:23:58:31]	"GET/Access/chapter1/s2-29.htmlHTTP/1.0"	200	3407

Figure 3- A Part of EPA log file after cleaning and filtering

3.4.2. PATTERN DISCOVERY (CASE-BASED REASONING METHOD)

The second phase of the process of the web usage mining is the pattern discovery. Finding patterns is a key component in the web exploring. This phase integrates algorithms and techniques from several research areas such as data mining, machine learning, and statistical pattern recognition.

Recommendation systems used data mining techniques (i.e., association rules, sequential pattern discovery, clustering and classification) to build and make recommendations using the knowledge obtained from the users' behavior and characteristics. Simply, Recommender systems try to guess the user's way of thinking, and to identify his best and closest interests and recommend his preferences to him/her. The CBR is known as a method which models the human behavior in dealing with new problems. Thus, the experiences gained in solving previous problems are exploited as a guide for solving new problems. The CBR cycle is consists of four stages, namely, retrieval, reuse, revising, and retaining. Firstly, in the login of a new issue, the CBR retrieves the conditions of the Case-base case that is most similar to the problem. In the second stage, the solution retrieved is re-used. In the third step, in order to fit a new problem, the solution is revised. In the fourth step, the solution is revised, and saved and reused for the future. The criteria for the similarity between the current user and the previous user are determined by comparing the Cases and using the Peer to Peer similarity of the Cases, clustering the Cases and, if necessary, the association rule mining in relation to any Cases and are achieved via the RapidMiner software within each case.

3.4.2.1. CASE ORGANIZATION

The Case of the proposed model includes the user IP address, number of pages viewed by that user, and the numbers specified to the pages viewed by users. However, the items are listed in the form of a matrix, and are applied to all users. So, every record will be considered equivalent to a Case.

3.4.2.2. RETRIEVAL PHASE

In the login of a new issue, the CBR retrieves the case that has the most similarity to the new issue from the Base Case. In the retrieval phase, the new case is compared to the cases of the Case Base, and the

most similar case is extracted. To compute the degree of the similarity, some criteria that can properly assess the similarity are required. So far, many criteria such as cosine similarity measures etc. have been proposed. In this study, in order to measure the similarity in cases, the Case-based reasoning method utilized the Exact Similarity, and so as to improve results, Clustering is used.

To calculate the similarity between two samples in the CBR method, in the Exact Similarity method, those cases are required to be retrieved from the Case Base that their number of pages observed is greater than or equals to the number of pages viewed by the Test Case in order to provide the desired solutions to the Test Case. After doing this, via similar criteria mentioned in (1), the similarity of the mentioned Test Case is assessed regarding all of the samples in the Case Base.

$$Sim_{local}(t.f, c.f) = \begin{cases} 1 & c.f = t.f \\ 0 & c.f \neq t.f \end{cases} \qquad (1)$$

If the two features of peer to peer cases are the same, the value is equal to "1"; otherwise the value "0" is assigned to the similarity of the characteristics of our cases. After doing this for each feature of all the cases, the values obtained for each case are added and thus the similarity of the Test Case to the total Case is obtained. Also, the explained process is used as a sliding window on all the features of the training data. Therefore, it is possible to select the highest existing similarity, and recommend the available solutions found in those Cases (the next page scrolling by the user) to the specified Test Case (the user while navigating).

3.4.2.3. REUSE PHASE

In the second step of the CBR process, the solution of the most similar retrieved case is copied and sent for the next step.

3.4.2.4. REVISE PHASE

In this step of the CBR, the feedback obtained from the degree of similarity of the retrieval phase, the solutions found in the cases with a high degree of similarity are investigated, and propounded as a set of proposed solutions to the Test data.

3.4.2.5. MAINTENANCE PHASE

Finally, the revised case (proposed solutions) is compared with the cases in the case base. If there are not any cases similar to this case in the case base and the proposed solution is appropriate, then the revised case is stored and added to the case base. Due to this process, the knowledge of the system is increased.

One of the main problems in the scope of recommendation systems is that at the beginning of the work, for some time the system is unable to provide an appropriate response, because the knowledge base is built based on a limited number of cases. For many of the issues presented, with the passage of time, the frequent use of the users, and the maintenance phase of the system, y the knowledge bases grows using the users' information, but in this period of time , the system performance may not be at an acceptable level. Until now, few approaches, such as hybrid approaches, have been proposed to solve

this problem. Hybrid recommendation systems are those systems in which the combination of one or more algorithms is used in order to achieve the maximum performance. A wide range of studies conducted in this area declare that hybrid systems are successful systems, and the main argument is focused on the method selection according to the conditions and characteristics of the problem of combining algorithms [26]. Thus, to reduce problems caused by slow start and to increase the efficiency of the proposed system during the startup, Apriori algorithm has been applied. In other words, if there is no similarity between the new case and the previous cases, association rule mining is used to make recommendation to the new case. For this purpose, we will gain the association rule mining associated to any case is measured by the RapidMiner software in each case.

3.4.2.6. ASSOCIATION RULE MINING IN RAPIDMINER

The next data mining method used in the pattern discovery phase of this project is the association rule mining. Association rules show the interactions among items in terms of their occurrence patterns in transactions with each other (regardless of their order) [27]. In the case of Web transactions, association rules show the relationships of the page views based on the user circulation patterns. Most of the association rule discovery approaches can be based on the Apriori algorithm. In this section, an ARFF file which contains all of the user navigational pages, is generated to be the input to the data mining RapidMiner software in order to perform the association rule mining based on the previous steps via the Matlab software. The process of the association rules is shown in Figure 4. In Figure 4, four blocks, namely, "Read ARFF file", "Preprocessing block "," Fp-Growth block", and "generate the association rules" are indicated. The preprocessing block is used to separate the file content. Such rules and the results can be used to optimize the structure of the website.

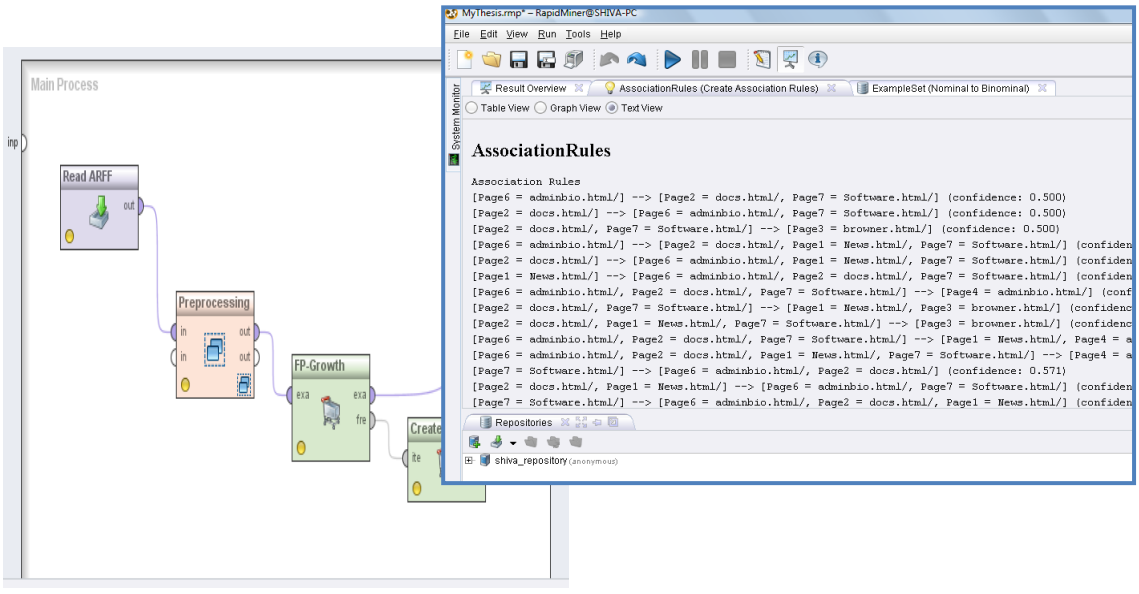


Figure 4 – The association rule mining process in RapidMiner and the results

3.4.3. RECOMMENDATION PHASE

Finally, the goal of this phase is to provide a private link. For the selection of training and testing sets, K-Fold is used. The advantage of K-Fold is that via the repetition of the random sub-sample, the whole set is used as the training data and the test data. In order to prepare the testing and training data, the

value of parameter K in K-Fold method is considered to be 10, and the abovementioned processes are performed on the samples.Table1 shows the percentage of the incorrect recommendations made using Exact Similarity and the Apriori Algorithm, after a maintenance phase data.

Table1-The percentage of the incorrect recommendations made via Exact Similarity and Apriori Algorithm

	K-Fold No. Number Of Error =%									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Exact Similarity	**88**	**88**	**83**	**85**	**88**	**87**	**15**	**7**	**3**	**3**
Exact Similarity+Apriori	**85**	**85**	**80**	**82**	**85**	**87**	**6**	**3**	**2**	**3**

3.5. MSNBC LOG FILE

To further improve the results and use the web content mining, as well as to integrate the Web content mining (WCM) and Web usage mining(WUM) using the case-based reasoning, the page content navigation EPA log file is used . But, due to the fact that the standard log files were gathered several years ago, and due to the lack of those pages or the change in the name of the pages, we could not view and analyze the words on them, and therefore, could not achieve the content. Also, owing to some security issues and to having the standard logs, it was impossible to access some logs. .Therefore, we tried to use a log with the specified case pages. In this study, MSNBC log file is used, which has 17 subjects categories. The subject categories include: news, music, weather, health & medical, sports, lifestyle and so on. This log file has 989,818 users and the average number of pages viewed per user is 5.7. Also, it can be downloaded from the website on http://archive.ics.uci.edu/ml/machine-learning-databases/msnbc-mld/msnbc.data.html.

3.5.1. PREPROCESSING

According to the EPA log, we implement all the work done in the MSNBC log file, except that the operations that can be performed in the data preparation phase is much less than those in the EPA's log file. At first, MSNBC text file converted into the form of a matrix, and then the number of pages for each user is calculated and put it in the records of that user. As mentioned in the previous section, the elements in each case are the IP address, the number of pages viewed by each user, and the number of topics specified to the pages viewed by the users.

3.5.2. CBR MODELLING

Similar to the procedures performed on the EPA log in the pattern discovery phase, we implement the Exact Similarity method on the MSNBC log file. In addition, because of nominal data, Clustering Method is used to for improve the results on the MSNBC log file.

- **Clustering Nominal Data:** The goal of clustering is to divide the data into several groups, and in this classification, the data of the different groups have to be as different as possible, and the data in the same group should be very similar. So far, different clustering algorithms for large databases are presented. But, all clustering methods are done based on distance functions which compute the amount of similarities between different items. For example, in the K-means clustering, the distance function of the Euclidean distance is used. But, in this study, due to Nominal data, the mentioned criteria cannot be used. So, here, a clustering method is used for the Nominal data [28]. By converting the original data into binary values, the clustering algorithm was run on the preprocessed data set according to the method described in [28]. Figure 5 shows the log file as a binary system. Finally, after clustering, in order to determine to which one of the existing clusters the test case is the most similar, the values of characteristics of the test case are compared with the center of each

cluster, and its distances to the center of all other clusters are measured. After this stage, the minimum distance is chosen and introduced as the desired cluster. Then, a number of existing solutions (via tests and evaluation) in the cluster is offered to the Test Case. It should be noted that like the association rule mining related to the EPA log file, the mentioned processes on the MSNBC log is also implemented and used to improve the results.

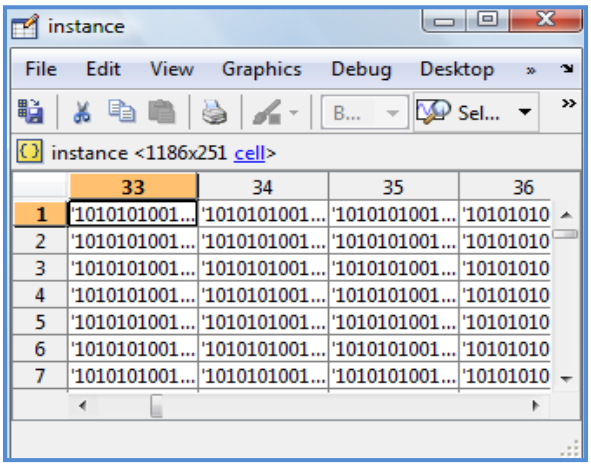


Figure 5 - A part of MSNBC log file as binary

3.5.3. RECOMMENDATION PHASE

The recommended phase solutions obtained are sent to the user platform. Finally, in the maintenance phase, the proposed solutions are compared with samples from the case base. If there is not any similar sample in the case base and the solution is appropriate, then the revised case is stored in the case base and added to it. Accordingly, due to the process, the knowledge of the system is added. Table2 shows the percentage of the incorrect recommendations made using Exact Similarity and Clustering Nominal data along with Apriori Algorithm in the recommendation phase.

Table2-The percentage of the incorrect recommendations made via Exact Similarity and Clustering in the recommendation phase

	K-Fold No. Number Of Error =%									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Exact Similarity	**51**	**50**	**47**	**47**	**45**	**54**	**55**	**49**	**50**	**38**
Clustering +Apriori Algorithm	**18**	**33**	**36**	**25**	**33**	**35**	**27**	**36**	**31**	**25**

4. EXPERIMENTAL RESULT ANALYSIS

Based on the results of the whole examination system, for more complete understanding of the performance of the proposed system, the Precision parameter is used as a benchmark to evaluate the work. Precision is a parameter that represents how accurate the proposed URL is and how much it is consistent with the actual behavior of the users. Finally, to create a single estimate, the mean of the Ks obtained from the accuracy rate of the proposed system is taken.

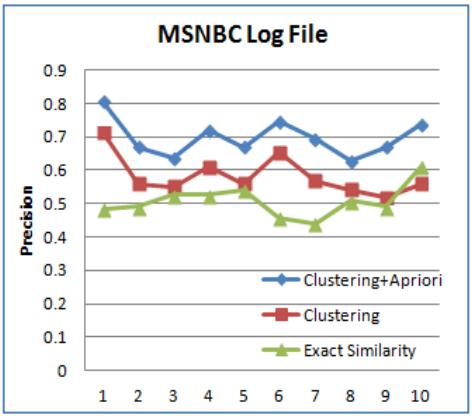


Figure 7 – The comparison of the accuracy of the MSNBC log files using Exact Similarity and Clustering with Apriori Algorithm

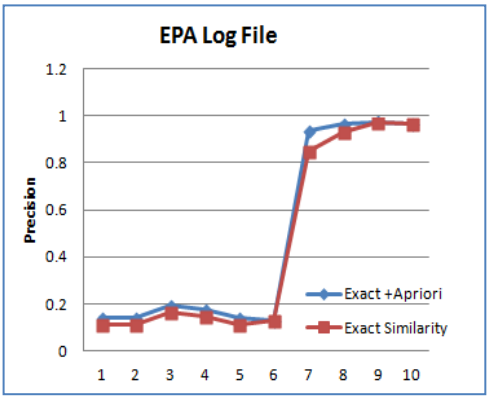


Figure 6 – The comparison of the accuracy of the EPA log file using the Exact Similarity and Apriori

According to Figure 6, it can be concluded that the EPA log files in Exact Similarity approach, the accuracy was 44%, and along with using the Apriori algorithm, it was improved to 48%. While the rate in the MSNBC log file (Figure7) was 51% in the Exact Similarity approach and 59% in the Clustering approach, and via utilizing a Apriori algorithm, the accuracy is improved to 71%. Figures 6 and 7 illustrate the results. Our evaluation shows that the proposed method along with the standard case-classified logs can provide more effective recommendations on the navigation of the next page to the user in comparison with the logs without any case classification.

4.1. THE ACTIVATION OF THE CBR ON THE PROPOSED SYSTEM

The most important achievement of the study is the introduction of a new system for recommendation subsequent user navigation pages using user records and behaviors in web usage mining and case-based reasoning. In this regard, other achievements include:

- Case-based reasoning techniques have been used to solve duplicate problems and it has been concluded that applying CBR and using previous experiences have been effective in reducing the number of iterations required and improving the results obtained from the implementation of the algorithms. It's a bargain at times.
- Checking the retrieval accuracy of the cases means that if one case is called from the case database, the system must deliver the same case with 100% similarity criteria. For this purpose, a number of cases were selected from the case library and tested. The test results show that the average similarity for the experimental cases in the case database is 100%. Therefore, it can be concluded that the accuracy of the system recovery is appropriate in bidding for the next user navigation page.
- A new approach used is the Nominal Web Data Clustering algorithm for clustering cases, which improves the results compared to the Exact Similarity method.
- Another issue highlighted in this study is the problem of not providing recommendations to some users who are not similar to users in the log file. It is also one of the challenges facing recommender systems. To address this problem, the Apriori algorithm was designed separately to contribute recommendations. The results of the nationwide review of the system presented in section 3 show that when recommendations are considered inappropriate and inadequate, the use of association rules increases the overall efficiency of the recommender system.

6. DISCUSSION

In this study, a new predictive method was presented using CBR method and web usage mining, which offers the user a choice of the next page to navigate. For example, if the user has seen page A and then page B and page C, we will recommend page D to the user in the tracking sequence, as well as the recommendations provided by previous users and checking their records and how they are navigated. This means that previous users followed page D after pages ABC, so the current user will most likely also navigate page D, and the similarity criterion of the current user with previous users will be get by comparing the cases using the sequential pattern inside each case. The Comparison of the relevant studies with this study in various fields along with their advantages, disadvantages, and evaluation metrics is provided in Table 3. However, there is a dearth of research on CBR technique in the field of web usage mining and the analysis of log file. In the current study, which is based on the reuse of previous work experience, the Case-Based Reasoning technique, is utilized efficient solutions in this regard.

Table 3.The Comparison of the relevant studies with current study in various fields

	Reference #	**Dataset**	**Advantage**	**Disadvantage**	**Evaluation Metrics**		
					Algorithm	CBR	**Measure name**
Current study	-	EPA and MSNBC log File	enough recommendations (Applied RapidMiner and Matlab)	enough recommendation	association rule mining +Appriori+Nominal	✓	**Cosine and Exact similarity**
Identify navigational pattern of web users	[29]	EPA log File	Applied RapidMiner	No enough recommendation	association rule mining	×	-
Sequence alignment and homology search with blast and clustalw	[30]	NCBI database	protein sequences	-	Sequence algorithm	×	-
Recommendations For Academic Major Of Students	[31]	students opinions	enough recommendation	No enough recommendation	Apriori	✓	**Exact similarity**

6. CONCLUSION AND FUTURE WORK

In recent years, the CBR has indicated high capabilities in various fields such as decision making, prediction, fault diagnosis, planning, information retrieval, etc. The main goal of this research is to design and implement a system which can recommend the navigational pages to the users by combining Web Usage Mining and Case-Based Reasoning and utilizing Association Rule Mining. In order to ensure the proper functioning of the proposed system, two sets of the standard data such as MSNBC and EPA log file used in this article. In this regard, a variety of tests and studies were conducted to evaluate the system. The results indicate that almost 70% of the recommendations of the proposed system to the testing data are accurate. One of the issues considered in this study is that there are no recommendations to those who are different from the existing users in the log file. Also, it is one of the challenges facing the recommendations systems. To deal with this problem, the Apriori algorithm is separately designed to participate in making recommendations. The results of the overall system presented in Section 4 indicate that in cases where the recommendations are considered inadequate and improper, using the association rules increases the overall efficiency of the recommendation system. Also, the prediction process can be undertaken via other methods such Particle Filter, Hierarchical Method, Semi-Supervised Method, etc.

Compliance with ethical standards

Conflict of interest: The authors declare that they do not have any conflicts of interest.

REFERENCES

1. Maimon, O. and L. Rokach, *Data mining and knowledge discovery handbook.* 2005.
2. Srivastava, J., et al., *Web usage mining: Discovery and applications of usage patterns from web data.* Acm Sigkdd Explorations Newsletter, 2000. **1**(2): p. 12-23.
3. Raphaeli, O., A. Goldstein, and L. Fink, *Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach.* Electronic commerce research and applications, 2017. **26**: p. 1-12.
4. Wang, G., et al. *Unsupervised clickstream clustering for user behavior analysis*. in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016.
5. Neelima, G. and S. Rodda. *Predicting user behavior through sessions using the web log mining*. in *2016 International Conference on Advances in Human Machine Interaction (HMI)*. 2016. IEEE.
6. Gauch, S., et al., *User profiles for personalized information access*, in *The adaptive web*. 2007, Springer. p. 54-89.
7. Lopes, P. and B. Roy, *Dynamic recommendation system using web usage mining for ecommerce users.* Procedia Computer Science, 2015. **45**: p. 60-69.
8. Malik, Z.K. and C. Fyfe, *Review of web personalization.* Journal of Emerging Technologies in Web Intelligence, 2012. **4**(3): p. 285-296.
9. Wagh, R. and J. Patil, *Enhanced web personalization for improved browsing experience.* Advances in Computational Sciences and Technology, 2017. **10**(6): p. 1953-1968.
10. Romero, C., et al., *Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems.* Computers & Education, 2009. **53**(3): p. 828-840.
11. Sujatha, V., *Improved user navigation pattern prediction technique from web log data.* Procedia Engineering, 2012. **30**: p. 92-99.
12. Lee, C.-H., Y.-l. Lo, and Y.-H. Fu, *A novel prediction model based on hierarchical characteristic of web site.* Expert Systems with Applications, 2011. **38**(4): p. 3422-3430.
13. Tseng, V.S., K.W. Lin, and J.-C. Chang, *Prediction of user navigation patterns by mining the temporal web usage evolution.* Soft Computing, 2008. **12**(2): p. 157-163.
14. Varma, A. *ICARUS: design and deployment of a case-based reasoning system for locomotive diagnostics*. in *International Conference on Case-Based Reasoning*. 1999. Springer.
15. Montani, S., et al., *Diabetic patients management exploiting case-based reasoning techniques.* Computer Methods and Programs in Biomedicine, 2000. **62**(3): p. 205-218.

16. Schmidt, G., *Case-based reasoning for production scheduling.* International Journal of Production Economics, 1998. **56**: p. 537-546.
17. Shiu, S.C. and S.K. Pal, *Case-based reasoning: concepts, features and soft computing.* Applied Intelligence, 2004. **21**(3): p. 233-238.
18. Mullins, R., et al., *A Web Based Intelligent Training System for SMEs.* Electronic Journal of E-learning, 2007. **5**(1): p. 39-48.
19. Kolodner, J.L., *An introduction to case-based reasoning.* Artificial intelligence review, 1992. **6**(1): p. 3-34.
20. Eirinaki, M. and M. Vazirgiannis, *Web mining for web personalization.* ACM Transactions on Internet Technology (TOIT), 2003. **3**(1): p. 1-27.
21. Singh, B. and H.K. Singh. *Web data mining research: a survey*. in *2010 IEEE International Conference on Computational Intelligence and Computing Research*. 2010. IEEE.
22. Aamodt, A. and E. Plaza, *Case-based reasoning: Foundational issues, methodological variations, and system approaches.* AI communications, 1994. **7**(1): p. 39-59.
23. Changchien, S.W. and M.-C. Lin, *Design and implementation of a case-based reasoning system for marketing plans.* Expert systems with applications, 2005. **28**(1): p. 43-53.
24. Markov, Z. and D.T. Larose, *Data mining the Web: uncovering patterns in Web content, structure, and usage*. 2007: John Wiley & Sons.
25. Dinuca, C. and D. Ciobanu, *Improving the session identification using the mean time.* International Journal of Mathematical Models and Methods in Applied Sciences, 2012. **6**(2): p. 265-272.
26. Burke, R., *Hybrid recommender systems: Survey and experiments.* User modeling and user-adapted interaction, 2002. **12**(4): p. 331-370.
27. Raut, M.S.M. and D. Dakhane, *Comparative Study of Clustering and Association Method for Large Database in Time Domain.* International Journal, 2012. **2**(12).
28. Wang, B. *A new clustering algorithm on nominal data sets*. in *Proceedings of the International MultiConference of Engineers and Computer Scientists*. 2010.
29. Asadianfam, S. and M. Mohammadi, *Identify navigational patterns of web users.* International Journal Of Computer-Aided Technologies (Ijcax) Vol, 2014. **1**.
30. Hung, J.-H. and Z. Weng, *Sequence alignment and homology search with BLAST and ClustalW.* Cold Spring Harbor Protocols, 2016. **2016**(11): p. pdb. prot093088.
31. Asadianfam, S. and S. Asadianfam, *Recommendations For Academic Major Of Students With Case-Based Reasoning Method To The Case Of Iran.*