ISSN 2277-3061

WEB PERSONALIZATION WITH WEB USAGE MINING TECHNIQUES AND ASSOCIATION RULES

*Gelareh Kazeminouri[1]; Ali Harounabadi[2] and Seyed Javad Mirabedini[3]

[1]M.S in Computer Software, Department of Computer, Kish International Branch, Islamic Azad University, Kish Island, Iran

e.gelareh.kazemi@gmail.com

[2]Assistant Professor, Central Tehran branch, Islamic Azad University, Tehran, Iran

a.harounabadi@gmail.com

[3]Assistant Professor, Central Tehran branch, Islamic Azad University, Tehran, Iran

Jvd2205@yahoo.com

ABSTRACT

As amount of information and web development increase considerably, some technics and methods are required to allow efficient access to data and information extraction from them. Extracting useful pattern from worldwide networks that are referred to as web mining is considered as one of the main applications of data mining. The key challenges of web users are exploring websites for finding the relevant information by taking minimum time in an efficient manner. Discovering the hidden knowledge in the manner of interaction in the web is considered as one of the most important technics in web utilization mining. Information overload is one of the main problems in current web and for tackling this problem the web personalization systems are presented that adapts the content and services of a website with user's interests and browsing behavior. Today website personalization is turned into a popular event for web users and it plays a leading role in speed of access and providing users' desirable information. The objective of current article is extracting index based on users' behavior and web personalization using web mining technics based on utilization and association rules. In proposed methods the weighting criteria showing the extent of interest of users to the pages are expressed and a method is presented based on combination of association rules and clustering by perceptron neural network for web personalization. The proposed method simulation results suggest the improvement of precision and coverage criteria with respect to other compared methods.

Keywords

Web Mining; Web Personalization; Association Rules; Neural Network

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol. 15, No. 1

www.ijctonline.com, editorijctonline@gmail.com

1.INTRODUCTION

The advancement of the technology has smoothed the way for communicating with higher speed. Data-mining is extracting knowledge from great amount of data and is known as the most efficient technic in the course of knowledge detection. Web personalization is aimed to providing information required for users and has a leading role in improving the interaction of users with web. The objective of web proposing systems is directing users toward pages that meet their needs and interests in the best way. Worldwide web includes great deal of information and acts as a huge database. In the past the need to methods able to provide an efficient access to data and information extraction from them has been felt more than ever. However, today these needs are met by using main technics of data mining such as association rules mining, extraction of sequential patterns and clustering for extracting users conductive patterns and presenting suggestions based upon them (Carmona, 2012). Association rules have been used successfully in web recommender systems. Association rules mining is posed as a research ground in the field of data mining and along with detecting important association between informational articles in the huge database.

In this paper, we are set to develop a recommender system using web mining based on application and weighted association rules. Conflating the K-means clustering technics and weighted association rules are considered as the innovations of this study.

In section 2, the background materials regarding association rules and k-means clustering and perceptron neural network structure are presented. In section 3, number of conducted works in the debated field is examined. In the section 4, the proposal is debated. In this section k-means clustering and association rules are combined and the outcome is stated as the structure of a perceptron neural network. In Section 5, the implementation and evaluation of proposed method are examined. By measuring two criteria of precision and coverage, we evaluate the proposed method and finally in the section 6, the conclusion and future suggestions are presented.

1.1 Background materials

1.1.1 K-mean clustering algorithm

One of the valid methods of clustering is K-mean. We draw on K-means algorithm for clustering. Users' session based on visited pages similarity in each session are clustered. In this algorithm, k is obtained from the input and n existing object are divided into k cluster. In this method, similarity of each cluster with respect to objects average of that cluster is measured and it is considered as so-called the center of the cluster. K-means selects k objects (input sample) as center of the cluster randomly. Then it allocates other input samples based on minimum Euclidean distance from the determined center of clusters to appropriate clusters. We consider each vector as a cluster node (Bishnu, Partha , & Vandana, 2012). The criteria that should be minimized in K-means are stated as the relation (1):

$$E_{k-means} = \frac{1}{c}\sum_{k=1}^{c}\sum_{x\in Q_k}\left\| x - c_k \right\|^2 \qquad (1)$$

In relation (1), || || is the criteria of distance between points, c the number of clusters and c_k is the center of k^{th} cluster.

1.1.2 *Association Rule Mining*

Association rule mining (ARM) algorithm extracts the associations between items by finding items that already have appeared together in transaction of dataset D = $\{T_1, T_2, \dots, T_s\}$. If we consider a set $I = \{I_1, I_2, I_3, \dots, I_m\}$ each transaction T is a subset of I. the overview of association rule is as relation 2:

$$A \rightarrow B \quad [Support, confidence] \qquad (2)$$

The association rule $A \rightarrow B$ is measured by Confidence, support where A and B are two set item and a subset of I so that $A \cap B = \emptyset$.

Support of association rule is defined as a percentage of transactions that includes both of items of the set (Skillen, Prescott, & Caldwell Livermore, 2015).

The Confidence coefficient of association rule is conditional probability that is used between transactions in which they are placed. In other word, it states the degree of dependence of a certain goods to another. For examining the value and criteria of acceptability of association rules, we introduce two parameters including support and confidence coefficient of rules that the following relations shows how to calculate them:

Support coefficient $(A \rightarrow B) = Support\ (A \cup B) = P(A \cup B)$ (3)

$$Confidence\ coefficient(A \rightarrow B) = \frac{A \cup B}{A} \qquad (4)$$

(Support) law * (confidence coefficient) law = (quantity) law (5)

1.1.3 Neural network

Neural network is a method for calculating and is made based upon interconnection of several processing units (Romero , 2012). The network is constituted from desired number of neurons that associate the input set to the output one.

1.1.4 Perceptron neural network

One kind of neural network is made based on a computational unit called perceptron. A perceptron takes a vector from inputs with actual values and computes a linear combination from these inputs. If the result is greater than a threshold value, the perceptron output would be 1, otherwise it is -1. Perceptron is neuron with two-level motivational function which its weights and biases are upgraded by learning rules.

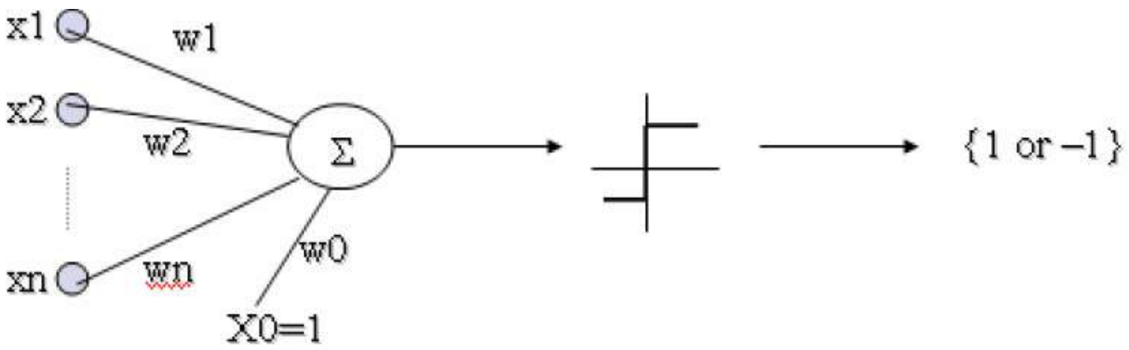

Figure 1: Perceptron neural network

Perceptron learning is finding correct values for weights (W). Hypothesis space H in perceptron learning is a set of all possible values for weight vectors. Each Boolean can show many Boolean functions such as NOR, NAND, OR, AND but it does not show XOR, because it is nor separable for linear functions and one cannot use perceptron network.

The output of perceptron can be obtained by relation (6):

O(X1, X2,…,PXn) = (6)

$$\begin{cases} 1 \text{ if } W0 + W1X1 + W2X2 + \cdots + WnXn > \\ -1 \;\; \text{Otherwise} \qquad\qquad\qquad\qquad\qquad 0 \end{cases}$$

1.1.5 Perceptron learning algorithm

1. We attribute random values to weights

2. We apply perceptron to training example. If an example is evaluated as incorrect, we correct the weight values of perceptron.

3. in response to the question of "are all training examples evaluated correctly? We take following steps:

- Yes → end of algorithm
- No →we go back to step 2.

In this paper we consider input p as users' feature that has been weighted by W and the perceptron output as clusters.

1.2 Literature review

In the paper of (Tyagi & Bharadwaj, 2012) a new recommender system is proposed by using quantitative association ruling for developing quantitative recommender for a new user. The method of proposed recommender s QARF (Filtering based QAR) and QARF/CF. (Liraki, 2014) has examined the web application mining using weighted association rules for extracting users' movement algorithm in web. In this paper, hybrid algorithm is presented from combination of fuzzy clustering method and weighted association ruling and fussy system and the users' sessions with similar surfing behaviors are clustered by c-means fuzzy clustering algorithm. Then, weighted association rules have been extracted from each cluster by using Apriori algorithm. Paper of (Jafari, Soleymani Sabschi, & Jalili Irani, 2014) states the utilization of web mining technics for designing effective recommender systems. As for Web Utilization Miner (WUM), the first session of association rules that is for pages that often are used together in a server session are identified. Since usually such transaction of database contains great amount of data, the association rule detection technics attempts to simplify the exploration space with view to items support. In the paper of (Ni, Liao, Wang, & Ren, 2009), , the association rules and pages clustering are adopted for recommending pages to users. Using two algorithm K-means and Apriori was basis of this paper. In the paper of (Mobasher, 2007) a comprehensive discussion of web personalization process is presented. The proposed method sets forth the details of clustering pattern detection technics, association rule mining, extraction of order pattern and essential models in implementing joint data in web as an integrated part of a web personalization system. Over steps of data mining cycle a mass of activities and various technics are adopted including: preprocessing and information integration from different sources and technics pattern detection. In the paper of (Mary & Baburaj, 2013) the web pages personalization has been turned into a popular event for customizing the web environments. The objective of personalization systems is considered to be providing users' needs, based on this technology the services that users require are personalized according to interests and features of users without their explicit assertion. (Ying, Zhou, Han, & Zhu, 2013) Has presented a method based on user context and collaborative filtering with aim of predicting the next demand of the user from web utilization pages. In this article for recommending pages preferred by the user a text factor is adopted for matching web pages and then by Merge-sort algorithm two candid proposed sets are merged. At first, the users are clustered considering the text factor and subsequently each cluster are assigned by scores so that the most similar page would place in the proposal set list.

2.DISCUSSION

2.1 Proposed algorithm

The proposed method in this paper is performed in both offline and online phases.

The steps of proposed algorithm

1. Preprocessing web server registrations for extracting users' sessions. Preprocessing includes steps such as: data cleaning, identifying users and sessions.

2. Converting sessions into sessions and allocating weight to visited pages at the sessions

3. Applying K-means clustering method for clustering similar sessions

4. Using perceptron neural network for expressing the outcomes (inputs are users' features and outputs are clusters).

5. Extracting weighted association rules from each cluster

2.2 Determining the pages weights

In this section, we present a criterion for page frequency and page observation duration for allocation of weight to page within users' session (Chen, Ze-Shui, & Mei-mei, 2014). If set P is set of pages accessible by users, P ={$p_1, p_2, \dots, p_m$} that each page is available with a unique URL and T={$t_1, t_2, \dots, t_n$} is a set of users transactions in preprocessed file of event record in which $t_i \in T$ is a subset of P pages. Each transaction of t_i is modeled as an m-fold vector of pages. $t_i = \{(p_1, w_1), (p_2, w_2), \dots, (p_m, w_m)\}$, in this relation w_i is the weight of page p_i in the transaction t_i.

I. page frequency: as the frequency of user visit to a page in a session is greater, that page is more important than other pages of the session. Page frequency is stated as following relation:

$$f\ (p) = \frac{f_{total}(p)}{l_s(p)+l_i(p)} \qquad (7)$$

Where, $f_{total}(p)$: total number of user's visits to a page

$l_i(p), l_s(p)$: Number of page external and internal links

II. **Page observation time duration:** is an essential criterion in determining the extent of user interest to the page and its importance. Time of observing the page is proportionate to the page size. Therefore if the length of the page is less, the observation time is less too. The page observation time duration is stated by following relation:

$$d\ (p) = \frac{d_{total}(p)}{length(p)} \qquad (8)$$

$d_{total}(p)$: Total observation time duration for a page

Length (p): page size in terms of bite

By combining two above criteria, the overall importance of the page is obtained. Here we use harmonic average for page weighting.

$$W(p) = \frac{2*f(p)*d(p)}{f(p)+d(p)} \qquad (9)$$

A user's session is a set of visited pages by the user during a visit from website.

$$S = < p_1, p_2, \dots, p_n > \qquad (10)$$

2.2.1 Extracting weighted association rules from each cluster

In this section, a set of weighted association rules are extract from each cluster including users' surfing information with similar behaviors and interests. For improving the process of personalization, the association rules mining allows us to attribute different weights to available items in the transactions. In this model, the greater weights suggest more important items. Most of association rules detection approaches are based on Apriori algorithm. This algorithm finds the observations of appeared pages that have had presence together frequently in many transactions and satisfy the support threshold determined by the user. In this study, we use Apriori algorithm for extracting association rules from each cluster. We extended this algorithm items to weighted items and we developed and examined it considering the degree of belonging of each item and the definitions related to it. Finally, we will use this model for recommending pages to users. Weighted association rules are expressed according to relation (11):

$$r = \langle (p_1, p_2, \dots, p_i), (q_{i+1}, q_{i+2}, \dots, q_{i+m}), (w_1, w_2, \dots, w_{i+m}), \alpha, \beta \rangle \in R \qquad (11)$$

In this relation, the former part is $(p_1, p_2, \dots, p_i)$ and the latter part of association rules is $(q_{i+1}, q_{i+2}, \dots, q_{i+m})$ and $(w_1, w_2, \dots, w_{i+m})$ is the corresponding weights with each one of the pages. We consider in this relation α and β as the weighted support coefficient and weighted confidence coefficient.

2.3 Web page recommendation window

For recommending pages in the most of previous approaches a sliding window with constant length is adopted on the user's current session based on which some suggestions are offered. The size of the window is the number of last pages observed by the user in the current session. This approach is not appropriate by itself, since in this approach the difference of importance of different pages for the user is not considered (Amatriain , 2011). Maybe a user makes backward or forward for finding its desired pages in the site and visit various pages. Therefore the best method is to consider the window size as changing.

In this paper, for recommending the pages to users based on weighted association rules, after improving the user's current session and allocating weight to observed pages by the current user, we select the appropriate length of observed pages sequence in current session that is used for contributing in pages recommendation. The pages that have not been interesting for the user and lack the information that the users require are eliminated from the user's current session and the pages recommendation is conducted based on user's interested pages. We draw on at the same time both "page weight" and" page location" in user's current session for selecting some pages from the session contributing in the page recommendation. After allocating the current user to the related cluster/s, the greatest degree of user belonging to the identified clusters is obtained and that cluster is selected as "objective cluster" (Ferrara, 2014). Pattern detection operation proceeds with using extracted association rules from that objective cluster.

The objective function in the following relation is defined so:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (12)$$

That || || is the criteria of distance between points and c_j is the center of j^{th} cluster.

For determining the degree of similarity of user's current session and association ruling, we consider the current session and extracted rules as m-fold vector. The degree of similarity of user's current session S $= \{s_1, s_2, \dots, s_m\}$, with the former part of association rules $r_l = \{w_1, w_2 \dots, w_m\}$ extracted from the objective cluster is calculated from following relation.

$$\text{Dissimilarity } (s, r_l) = \sum_{i:r_{li}} \left(\frac{2*(w(s_i)-w(r_{li}))}{(w(s_i)+w(r_{li}))}\right)^2 \qquad (13)$$

$$\text{Matchscore } (s, r_l) = 1- \frac{1}{4}\sqrt{\frac{\text{Dissimilarity}(s, r_l)}{\sum_{i:\, r_l>0} I}} \qquad (14)$$

Implementation and evaluation of proposed algorithm

In this method, we want to personalize web pages by web utilization mining technics and association rules. A website is consisted of great deal of web pages and each page can communicate to other pages through hypertext links.

In this research, we analyzed NASA site standard data, then extracted a set of required data and we examined the users' session information of 1995 in this site. After preprocessing for identifying users' sessions and session vectorization operation and allocating weight to visited pages in the session, the K-means clustering is presented for clustering similar session by using a perceptron neural network which its input is users and output is clusters. The results have been presented and finally the association rules of each cluster are extracted. Also, we draw on MATLAB software for implementing the proposed method.

Both precision and coverage are among effective parameters in system efficiency that can be calculated according to following relations.

The criteria of precision express the ratio of appropriate recommendations to the total number of recommendation. The objective is to find out how many recommendable pages are available for that user.

$$\text{Precision (rs,rp)} = \frac{|rs \cap rp|}{|rs|} \tag{15}$$

rs: a recommended set

rp: the pages observed by the user

The coverage criteria is the ratio of number of website recovered pages to total number of website pages that actually belong to user's session.

$$\text{Coverage (rs,rp)} = \frac{|rs \cap rp|}{|rp|} \tag{16}$$

Comparing the proposal with other methods

In this section, we compare the precision and coverage of proposed algorithm with other methods and the obtained results are shown in the figure 2 & 3. Two compared methods of (Liraki, 2014) examined the web utilization mining using weighted association rules for extracting users' surfing behavior in web that is named for brief (WAP) and the next method is related to (Singhal & Pandey, 2012) that has proposed the design of a website using association rules and clustering that is termed for brief (WAC). As it has been shown in following figures, the number of proposed pages increase, the precision declines and the coverage augments in all algorithms. The results suggest that the proposed algorithm features high precision and coverage with respect to other methods.

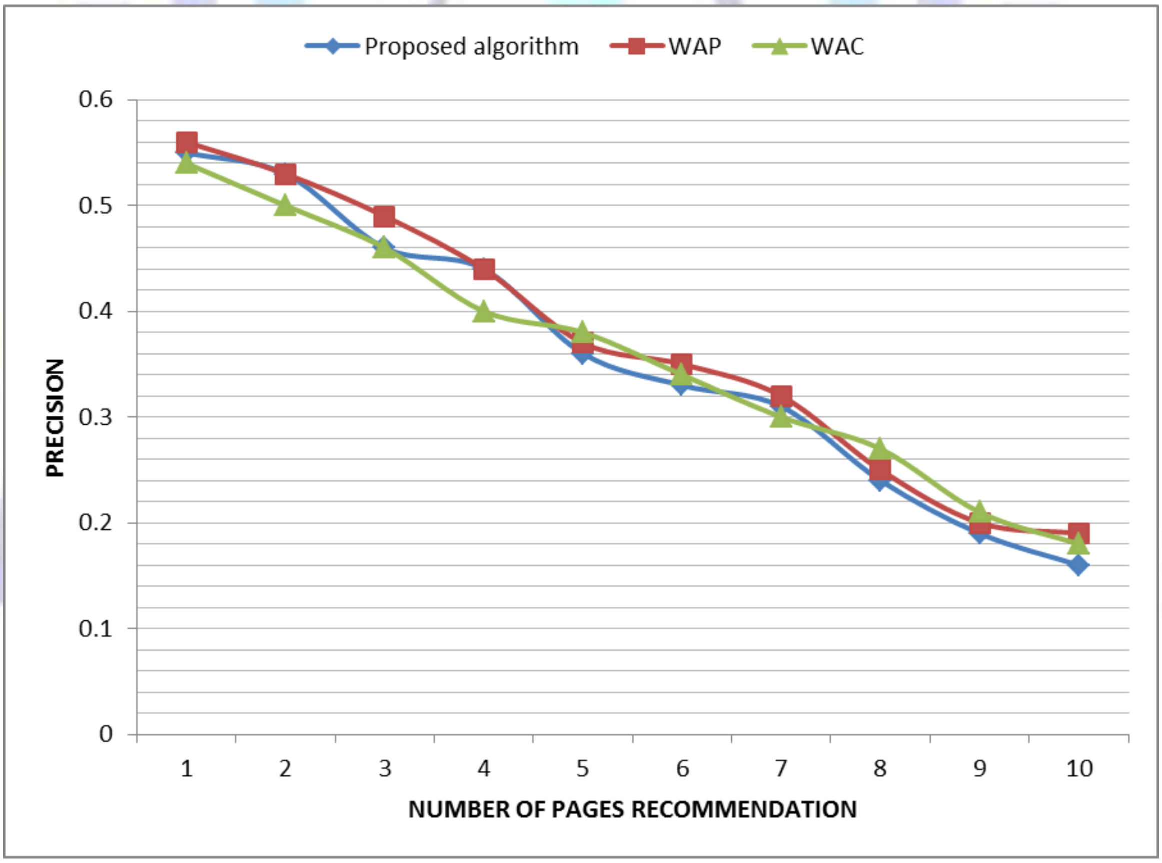

Figure 2: Comparing the precision of proposed algorithm with (Liraki, 2014)& (Singhal & Pandey, 2012) algorithm

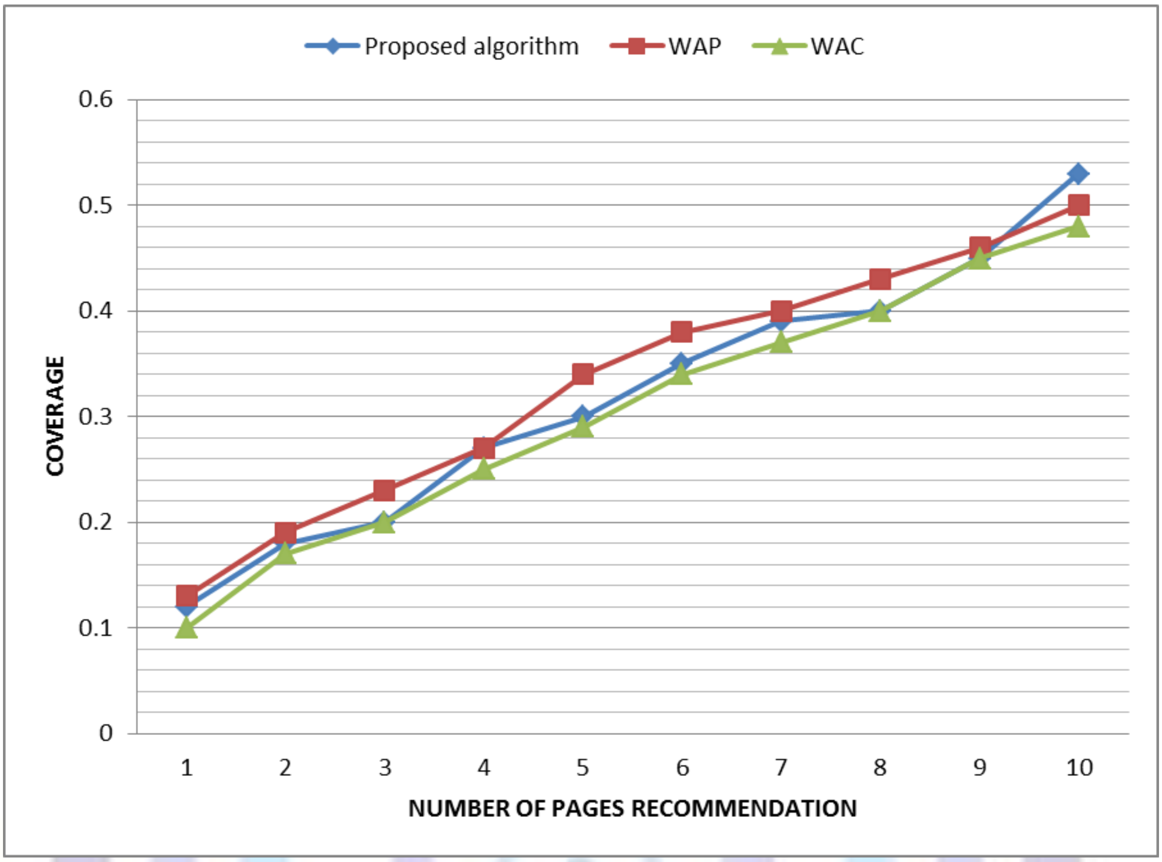

Figure 3: Comparing the coverage of proposed algorithm with (Liraki, 2014)& (Singhal & Pandey, 2012) algorithms

3. CONCLUSION

We briefly have examined the data-mining technics. One of the main parameters of web personalization system is user model that features high importance. As this model is developed better, the recommendations based on which are made are more precise and deeper. the main objective of this thesis includes extracting index based on users' behavior at web, presenting a clustering method so that users with different interests would be placed in different clusters and developing a recommendation system in the web, so that can be used in purposeful advertisement and the search engines and e-commerce (practical objective). In this paper, a method is presented for website personalization using utilization and association rules based web mining. The page observation time duration and page observation frequency criteria are used for allocating weight to pages. The weighting criteria that show the extent of interest of users to the pages are defined. In this research, by using web mining based on utilization and weighted association rules, a recommendation system has been developed. K-means clustering technics and association rules are conflated with each other, and the result is presented as a perceptron neural network structure. Also the MATLAB software has been used for implementing proposed method. The proposed method provides the opportunity of contribution of pages that are proportionate with interests and taste of users and are of high recommendation value in the set of recommendable pages. The analysis shows that the presented method features high precision and coverage percentage and is appropriate for web personalization.

3.1 Future suggestions

1. Adopting other clustering methods
2. Combining fuzzy methods, association and neural network
3. By using association rules and the proposed method in this paper, one can support further applications in this field in the future to use improved algorithms for developing association rules.

4. REFERENCES

[1] Carmona, C. J. (2012). Web usage mining to improve the design of a e-commerce website:. OroliveSur.com." Expert Systems with Applications , 11243-11249.

[2] Bishnu, Partha , S., & Vandana, B. (2012). Software fault prediction using quad tree-based K-means clustering algorithm. Knowledge and Data Engineering, IEEE transactions on 24.6, 1146-1150.

[3] Skillen, Prescott, R., & Caldwell Livermore, F. (2015). Associative search engine. U.S. Patent no.

[4] Romero , C. (2012). Web usage mining for predicting final marks of students that use Moodle courses. Computer Applications in Engineering Education, 135-146.

[5] Tyagi, S., & Bharadwaj, K. (2012). Enhanced New User Recommendations based on Quantitative association Rule mining. ELSEVIER the 3rd international conference on Ambient Systems, Networks and Technologies, 10, 102-109.

[6] Liraki, Z., Harounabadi, A., & Mirabedini, J. (2014). application of web mining using weighted association ruling for extracting users browsing patterns in the web. Ninth symposium of advancements of science and technology, institute of high education of Mashhad Khavaran, (pp. 1-9).

[7] Jafari, M., Soleymani Sabschi, F., & Jalili Irani, A. (2014). applying web usage Mining Technique to design effective Web recommendation systems: A case study. ACSIJ Advances in Computer Science: an International Journal, 3(2), 78-90.

[8] Ni, P., Liao, J., Wang, C., & Ren, K. (2009). Web information Recommendation Based on User Behaviors. IEEE 2009 World Congress on Computer science and Information engineering, Vol. 4, Date of Conference, (pp. 426-430). Los Angeles.

[9] Mobasher, B. (2007). Data Mining for Web Personalization, P. Brusilovsky, A. Kobsa, and W. Nejdl (EDS.): The adaptive Web, LNCS 4321. Springer-verlag Berlin Heidelberg, 90-135.

[10] Mary, P., & Baburaj, E. (2013). Constraint informative rules for gentic Algorithm-based Web Page reccomendation system. Journal of computer Science 9 911), Published Online 9 (11) (Http://www.thescipub.com/jcs.toc), ISSN. 1549-3636, 1589-1601.

[11] Ying, Z., Zhou, Z., Han, F., & Zhu, G. (2013). Research on Personalized Web Page Recommendation Algorithm Based on User Context and Collaborative Filtering. Software Engineering and Service Science (ICSESS), 4th IEEE International Conference, 220-224.

[12] Chen, N., Ze-Shui, X., & Mei-mei, X. (2014). Hierarchical hesitant fuzzy K-means clustering algorithm. Applied Mathematics- A Journal of Chinese Universities 29.1, 1-17.

[13] Amatriain , X. (2011). Data mining methods for recommender systems. Recommender systems Handbook. Springer US, 39-71.

[14] Ferrara, E. (2014). Web data extraction, applications and techniques: A survey. Knowledge-Based Systems 70, 301-323.

[15] Singhal, V., & Pandey, G. (2012). A web based Recommendations based on Quantitative Association Rule mining. ELSEVIER the 3rd International Conference on Ambient Systems, Netwroks and Technologies, 10, 102-109.