A Recommender System for the Web: Using User Profiles and Machine Learning Methods

Samane Rajabi
Department of computer engineering, islamic azad university
Central Tehran branch, Tehran, Iran

Ali Harounabadi
Department of computer engineering, islamic azad university
Central Tehran branch, Tehran, Iran

Vahe Aghazarian
Department of computer engineering, islamic azad university
Central Tehran branch, Tehran, Iran

ABSTRACT
Web development without an integrated structure makes lots of difficulties for users. Web personalization systems are presented to make the website compatible with interest of users in both aspects of contents and services. In this paper extracting user navigation patterns is used to capture similar behaviors of users in order to increase the quality of recommendations. Based on patterns extracted from the same user navigation, recommendations are provided to the user to make it easier to navigate. Recently, web browsing techniques have been widely used for personalization. In this study, a method is proposed to create a user profile with the web usage mining by clustering and neural networks in order to predict the user's future requests and then generate a list of the pages of user's favorites. Simulation results shows that proposed method will increase the accuracy of recommender systems.

Keywords
User profile, neural network, clustering, web usage mining.

1. INTRODUCTION
Continuous growth in size and usage of the World Wide Web creates new methods of design and development of online information services. Most web structures are large and complex, and users often do not achieve their goals in their research, or receive ambiguous results. On the other hand, the e-business sector is evolving rapidly and requires the web to predict customers' needs more than before. Web personalization is the prerequisite for improving the usability of web pages. In fact, personalization systems satisfy the user needs without expressing them explicitly. Web browsing techniques based on the data type are divided in three categories of web content mining, web structure mining, and web usage mining [1]. Recently, web mining techniques are widely used for personalization; the web usage mining techniques are used for the detection of navigation patterns of the users and the information in the web server log files [2]. Many researches were conducted in this field, mainly based on the information of user behaviors in interaction with a web page to extract hidden knowledge and navigation patterns.

In this paper the web usage mining and neural networks are implemented in order to predict user future requests and then generating a list of favorite pages for users. In this study various user interactions on the web are tracked and clusters are created based on user interests. By the aid of neural networks and clustering navigation patterns are created, in order to predict user future requests and tabulating a list of user's favorite pages. Pre-processing the log file and generating user profile according to the user session and visited pages is the primary step. Creating profile can be done based on different parameters, in this article Frequency, Duration and Date are used as parameters. The accuracy of applied algorithms increases by using these parameters to find the exact interests of a user. After getting user preferences and creating user profiles, K-means algorithm of clustering method is used to obtain navigation patterns. After clustering, the recommender's engine which works by neural network is implemented. The neural network system's input is navigation pattern and the output is more similar cluster to user session, to predict the user's next move.

The rest of the paper is organized as follows: in section two some examples are reviewed, and in section three some concepts like clustering algorithm is explained, the fourth section describes the purposed method, and the fifth section describe the simulation results.

2. RELATED WORK
This section reviews the work carried out in the context of Web personalization and create a profile for users.

Zhong and authors [3] presented a method which the user profile is made up of three stages. The first step is to identify the useful data. In second step, K-means algorithm is used for user session clustering. The results show that this method is an effective role to improve the user model compared to the previous methods. Maheswari, B [4] used the theory of distribution in Dempster-Shafer's theory, the belief function similarity measure in this algorithm adds to the clustering task the ability to capture the uncertainty among Web user's navigation performance. And experiments about the accomplishment of preprocessing and clustering of web log the experimental result shows the considerable performance of the proposed algorithm. Azimpour, M. [5] introduced a measure of similarity to compare the user's session. Web session clustering strongly depends on the similarity measure. The observed results show that the proposed method is very effective in recording user session characteristics. Valera, M. [6] proposed an efficient sequential pattern mining algorithm to identify frequent sequential web access patterns. They believe that the aim of discovering frequent sequential access patterns in web log data is to obtain information about the navigational behavior of the users. The access patterns are retrieved from a Graph, which is then used for matching and generating web links for recommendations. Mobasher and Nakagawa [7] presented techniques for web personalization based on association rule. Murat G. and authors [8] review four different combination methods and suggest methods to correct each of them. Also they did comparative evaluation on these four different methods that show how various

techniques and hybrid method proposed is effective in prediction accuracy. A recommender motor provided that the combined results of several recommender techniques based on web usage mining. The proposed hybrid approach, combining the results of different techniques and a set of proposal for a new user is generated. The authors have proved that the proposed system has higher accuracy than the hybrid systems.

3. THE PRELIMINARY REQUIREMENTS AND DEFINITIONS

This section briefly describes the concepts needed in the proposed method.

3.1 Web Personalization

Web personalization defines as any operation that comply website services or information with the requirements of a particular user or group of users. This is achieved by using knowledge of user observations and personal interests, combined with the content and the website structure. The ultimate purpose of a web personalization system is to provide information of users interest without expecting the explicitly request [1].

3.2 Web Usage Mining

Web usage mining is a kind of web mining techniques, including automatic discovery of user access patterns from one or more web servers. Organizations often produce massive amounts of data every day. More information will be created automatically by web servers and stored in the server access logs [2]. In other words, web usage mining focused on secondary data from user interaction instead of working on main web data (structure and content) and apply data mining techniques on such data so that discover interesting usage patterns.

3.3 Clustering

The purpose of clustering is to group data in clusters so that the similarity among all members of the cluster is maximized, while the similarity between members of different clusters is minimized. In [9] defined clustering as:

Given a set of input patterns $X=\{x_1, x_2,\ldots,x_n\}$, include n objects where each one is from the collection of equal size vector with the length s in terms of properties. These objects must be clustered in K groups named $C= \{C_1, C_2, \ldots, C_k\}$, which do not overlap with each other. So that the following three conditions are shown by the following relations, are formed:

$C_i \neq \emptyset$, i=1,…, K

$C_1 \cup C_2 \cup \ldots \cup C_K = X$ (1)

$C_i \cap C_j = \emptyset$, I, j =1, …, K and $i \neq j$

3.3.1 K-Means

This algorithm was introduced in 1967 by MacQueen for the first time. The basis of method is very simple, k-means clustering aims to partition n object into k clusters in which each object belongs to the cluster with the nearest mean.

4. THE PERPOSED METHOD

In this section, the proposed method is presented in order to generate suggestions for user future requests. The method presented in this paper is based on web usage mining in web server logs. And an algorithm is presented for constructing navigation patterns and finally provide proper offer for users. The basis of the work is pre-processing the log file and then using web usage mining techniques to make the session's vector. After making the navigation patterns by using features of user behavior and clustering, recommender engine suggests a list of user's favorite pages. The proposed system consists of several stages (see Figure 1), each stage of the algorithm described in the following description.

4.1 Pre-processing

The preparation is usually complex and time consuming. Data exists in this section are as the web server logs. This pre-processing is performed to identify web access sessions. Pre-processing on the web server logs should be performed before applying web mining algorithms and same as [10] pre-process is split into three steps:

4.1.1 Data cleaning

This phase includes all operations that will delete the useless data. In fact at this stage of pre-processing the removed data is as follows:

- requests for image files
- failed requests, For instance, the requests have been faced with HTTP error
- request which response except Get and Put

4.1.2 User session identification

At this phase, the user sessions are detected from log files. Set of pages visited by a user during a specific website mining called user session. The sessions represent user behavior so that they are important in the process of pattern discovery. Various methods have been proposed to identify user sessions which categories in time oriented method and navigation oriented method. In this study time oriented methods are used. The intended pages as a session are pages in a period of less than or equal to a certain time period requested. In these algorithms, the time considered is 30 minutes.

4.1.3 Forming the data

This is the last step of preprocessing of data. The data must be in a particular format so web mining techniques can be applied on them. The data set can be stored in a relational database.

4.2 Create Profile and Session Victorization

User session defines weight of pages as a vector. Assume that P is total pages accessed by users so that $P= \{p_1, p_2 \ldots p_m\}$, Also S is the set of users access sessions $S= \{S_1, S_2 \ldots S_n\}$, where each $S_i \in S$ is a subset of P and each session S_i are shown with m-dimensional vectors as follows: $si=\{w(p_1, si), w(p_2, si), \ldots, w(p_m,s_i)\}$ Weight $w (p_i, s_j)$ must specifically identify the user interest in a web page. Weights can be determined in various ways. In this study, to determine the degree of user's interest Frequency, Duration and Date are used. [11] Investigated other ways to determine the weights. Frequency displays number of visits to a web page.

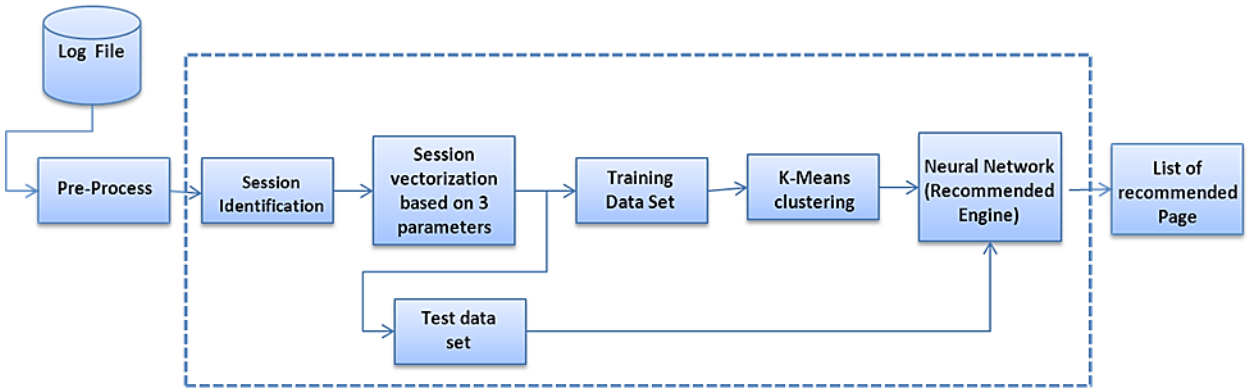

Fig 1: Purposed method

Frequency is calculated by the following formula:

$$frequency(page) = \frac{number\ of\ visits\ (pages)}{\sum_{page \in visited\ page} number\ of\ visits\ (pages)} \quad (2)$$

Duration, defined as the time spent on a page. If the user spends more time on a page, that page is more favored by the users and if a page is not favored by the users, the users have rejected that page quickly and will go to another page. However, length of page must also be considered. So the Duration is normalized by length of page. Duration is calculated by the following formula:

$$Duration(page) = \frac{total\ duration(page)/lenght(page)}{Max_{page \in visited\ page}(total\ duration(page)/lenght(page))} \quad (3)$$

More recent history is higher priority. To prioritize the pages for each day assigned a constant weight between zero and one, in this study considered NASA data set, for every day of this data set; dedicated a constant weight. Finally, user interest can be obtained by taking the harmonic mean of these three features.

$$Interest(page) = \frac{3 \times frequency(page) \times Duration(page) \times date(page)}{frequency(page) + Duration(page) + date(page)} \quad (4)$$

It should be noted that the amount of interest to be normalized between zero and one to be suitable for clustering.

4.3 Clustering

K-mean clustering algorithm is recommended. In this algorithm Manhattan distance is used to calculate the distance between two session vectors. After finding clusters consider a centroid vector for each clusters. Weight of every page in centroid vector is obtained by the average weight of pages in all of that cluster vectors. The output of this stage is navigation patterns of users. In other words, the centroid vectors of each cluster are the navigation patterns of users.

4.4 Recommender System Using Neural Network

This system received current user session and recommended a page for active users. To find the closest clusters to a user session, the neural network will be used. So by using navigation patterns, the network is trained. This means that each navigation pattern is considered as the input of neural network. The output of the network is number of cluster that has already been set for each navigation pattern. After training, the current user session must be prepared in a way that is suitable for input to the neural network. And this process is similar to session's vectorization. It is necessary to determine that this session belongs to which navigation pattern. For this purpose the current session profile, is given as input to the neural network and the appropriate cluster number for the session will determine. When the cluster number was determined pages of cluster which aren't in current session have the high potential to be the next page. And these pages will be recommended.

5. SIMULATION

In order to implement this method, NASA log file used like a lot of research work. As mentioned in section 4.1 preprocessing must be done in the initial stage. After removing the non-useful data for identifying session C # programming language is used. User sessions are detected with a threshold of 30 minutes. After this step, User session vectors are extracted. Sessions divides in two categories training and test. In order to implement K-Means algorithms Rapid miner software used. Neural network used in this study for recommender engine, three-layer network with an input layer, hidden, and output used in MATLAB software environment and hidden layer in the network has 50 neurons. The back-propagation network algorithm was used to learn from data and Levenberg Marquardt technique has been used as a learning technique.

The system trained by training data then simulation was performed by test data. Training phase is designated to assign a navigation pattern to each cluster by using two input matrices, target and navigation pattern. The results of the implemented system are presented in figure 2. For evaluating the proposed system, "association rule based recommendation systems" was used and the proposed system was validated with it [7].

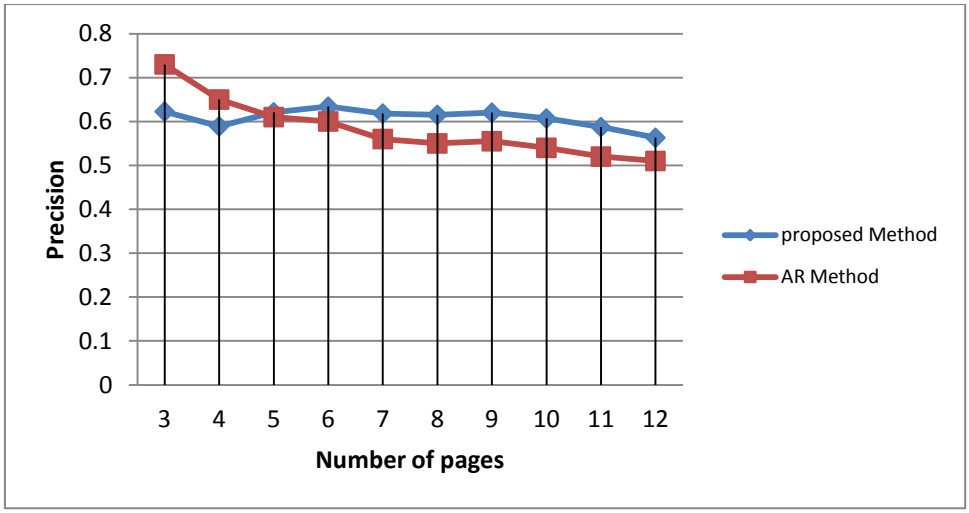

Fig2. Recommendation Precision Comparison with AR method

In this paper with the aid of Date feature, user behavior is tracked more accurately in order to enable the system to propose recommendations based on user's interests. Precision is the system's ability to generate accurate recommendations, in other words the precision of the system is determined by true recommendations ratio to all recommendations. As is shown in Figure 2 precision of the proposed system is better than association rule based recommendation systems after fourth recommended page. As expected the system precision increased with considering the Date feature.

6. CONCLUTION

In this paper various user interactions on the web are tracked and clusters are created based on user interests. By the aid of neural networks and clustering navigation patterns are created, in order to predict user future requests and tabulating a list of user interested pages. In this paper Frequency, Duration and Date are considered as parameters. Simulation result shows that the precision of used algorithms increased by using these parameters to find the exact interests of a user. After extracting user preferences and creating user profiles, using K-means algorithm of clustering method to obtain navigation patterns. After clustering and implementing the recommender engine by neural network, the user's next move is predicted. As expected the system precision increased with considering the Date feature. Result shows with increasing the number of recommended pages precision of the proposed system is better than association rule based recommendation systems.

7. ACKNOWLEDGMENT

The authors would like to give their special thanks to Dr Ali Hruonabadi and Dr. Vahe Aghazarian for their contributions on this work.

8. REFERENCES

[1] Pierrakos, D., Paliouras, G., Papatheodorou, CH., Spyropolous, C., (2003), "web usage mining as a tool for personalization: a survey", user modeling and user-adapted interaction13, pp: 311-372

[2] kumar malviya, R., malviya M.c., Soni, v.K., joshi, R., Purohit P., (2011) "survey of web usage mining" , international journal of computer science and technology, vol 2, issue 3.

[3] Zhong, J., and Li, X., (2010), "Unified Collaborative Filtering Model Based on Combination of Latent Features", Expert Systems with Applications, vol. 37, pp. 5666-5672.

[4] Maheswari, B., sumathi, P., (2014), "A New Clustering and Preprocessing for Web Log Mining", World Congress on Computing and Communication Technologies (WCCCT), DOI. 10.1109/WCCCT.2014.67,pp.25 – 29.

[5] Azimpour, M., Azmi, R., (2011), "A webpage similarity measure for web sessions clustering using sequence alignment", IEEE, International Symposium on Artificial Intelligence and Signal Processing (AISP), pp. 20 – 24.

[6] Valera, M., chauhan, U., (2013), "An efficient web recommender system based on approach of mining frequent sequential pattern from customized web log preprocessing", Forth International conference On Computing, Communications and Networking Technologies (ICCCNT), DOI. 10.1109/ICCCNT.2013.6726493, pp. 1- 6.

[7] Mobasher et al., B., Dai, H., Luo, T., & Nakagawa, M. (2001). Effective personalization based on association rule discovery from web usage data. In: Proceedings of the third ACM Workshop on Web Information and Data Management (WIDM01), held in conjunction with the International Conference on Information and Knowledge Management.

[8] Murat G., Sule G, (2010), "Combination of Web page recommender systems", Elsevier, Expert Systems with applications, vol. 37, no. 4, pp. 2911-2922.

[9] Xu R., (2005), "Survey of Clustering Algorithms", IEEE Transactions on Neural Network, vol. 16, no. 3, pp. 645–678.

[10] Tiedtke, T., Christian, M., Norbert, G., (2002), "AWUSA- A tool for automated website usability analysis" 9th international workshop on design, specification and verification of interactive sys.

[11] Mobasher, B., Cooley, R. & Srivastava, J., (2000), "Automatic Personalization Based on Web Usage Mining", in: Communications of the ACM, vol. 43, no. 8, pp. 142-151.