Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

Constructing a web recommender system using web usage mining and user's profiles

T. Mombeini[a]*, A. Harounabadi[b] and J. Rezaeian Sheshdeh[a]

[a]*Department of Computer Engineering, College of Computer, Ramhormoz Branch, Islamic Azad University, Ramhormoz, Iran*
[b]*Department of computer engineering, Islamic Azad University Central Tehran Branch, Tehran, Iran*

C H R O N I C L E

Article history:
Received June 12, 2014
Accepted 20 November 2014
Available online
November 24 2014

Web Personalization
Recommender System
Web Usage Mining
User Profiling
Fuzzy Clustering
Neural Network

A B S T R A C T

The World Wide Web is a great source of information, which is nowadays being widely used due to the availability of useful information changing, dynamically. However, the large number of webpages often confuses many users and it is hard for them to find information on their interests. Therefore, it is necessary to provide a system capable of guiding users towards their desired choices and services. Recommender systems search among a large collection of user interests and recommend those, which are likely to be favored the most by the user. Web usage mining was designed to function on web server records, which are included in user search results. Therefore, recommender servers use the web usage mining technique to predict users' browsing patterns and recommend those patterns in the form of a suggestion list. In this article, a recommender system based on web usage mining phases (online and offline) was proposed. In the offline phase, the first step is to analyze user access records to identify user sessions. Next, user profiles are built using data from server records based on the frequency of access to pages, the time spent by the user on each page and the date of page view. Date is of importance since it is more possible for users to request new pages more than old ones and old pages are less probable to be viewed, as users mostly look for new information. Following the creation of user profiles, users are categorized in clusters using the Fuzzy C-means clustering algorithm and S(c) criterion based on their similarities. In the online phase, a neural network is offered to identify the suggested model while online suggestions are generated using the suggestion module for the active user. Search engines analyze suggestion lists based on rate of user interest in pages and page rank and finally suggest appropriate pages to the active user. Experiments show that the proposed method of predicting user recent requested pages has more accuracy and cover than other methods.

1. Introduction

Today, due to the development of the web, electronic commerce, web services and web-based systems and the distinctive feature of the web (i.e. activity of its users), if a website cannot answer a user information request in a short time, the user will quickly and easily move on to other websites (Taghipour & Kardan, 2008). As predicting the information needs of clients is vital to every website, it has been the major concern of many organizations and researchers in recent years (Pierrakos et al., 2003). Usually, whenever a user is linked to a website, for each of his/her requests, one or more

*Corresponding author.
E-mail addresses: t.mombeini@0utlook.com (T. Mombeini)

doi: 10.5267/j.msl.2014.11.010

records of web servers are stored in history files. Multi-data analysis can be used to analyze users' behavior and performance. This process is usually called web usage mining (Mustapaşa et al., 2010). Web mining can be considered as the process of mining data on web content, structure and usage (Anand & Mobasher, 2003). The aim of web mining is to explore models and templates hidden inside web resources. The objective of web usage mining is also to explore web users' behavioral patterns. Exploring this vast amount of data created by web servers has different advantages (Nasraoui, et al., 2008). In recent years, web exploration techniques have been used as alternative strategies in web personalization as these techniques have reduced problems associated with general web filtration. Most web usage systems attempt to find better structure and clustering techniques, so that they could get access to a better model of users' navigation conditions. Data clustering is among the most common data mining techniques (Janssens, et al., 2009). Data processing is one of the most important indices in the world of information. Clustering is one of the best methods introduced for data processing. Clustering makes it possible to enter into the data space and identify the structure (Xu & Wunsch, 2005). Therefore, this mechanism is one of the most suitable ones to be used in the vast world of data.

The second section of this article describes the research literature and the third section discusses personalization architecture based on web usage mining. The fourth section describes the web recommender system. The fifth and sixth sections also include a case study and conclusion, respectively.

2. Related Works

During the past few years, there have been a large number of studies conducted on web recommender systems. Liu and Kešelj (2007) suggested an approach for classification of browsing patterns and prediction of users' future requests. The work started with an initial preprocessing of Web records and user sessions were extracted from the data set. Next, in order to identify user sessions, a vector of page weights was made. In order to calculate page weights, the following two criteria were used: frequency and duration of page view. Afterwards, the resulting sessions were clustered and the browsing patterns of users were obtained. The results were then combined with the content of related pages and profiles of browsing patterns were created. In this article, web page contents were obtained through extraction of n-gram characters. Following the extraction, classification of the browsing patterns and prediction of user future requests started.

Wu and Wu (2013) adjusted the membership and density functions and improved the conventional C-means clustering algorithm in order to solve problem in which the number of clusters used to determine the convergence of the objective function was inadequate. Next, personal preferences were divided into several groups in a way that users with similar preferences were put into the same group. Association rules of user preferences were identified and personalized knowledge was obtained. Afterwards, suggestions were provided to users through user review records and the extracted knowledge.

Almurtadha et al. (2011) introduced a recommender system to explore user priorities and suggested pages for future reviews. This system includes two phases. In the first phase, input data preprocessing and then K-means clustering algorithm were applied to the pages. Then a profile review was created for each cluster. In the second phase, first the active user profile was created based on previous sessions. Then the user profile was matched with the clusters obtained and the matching degree of active user profile for each cluster was calculated. Using the cosine coefficient, convenient offers were provided to the user based on the matching degrees and the degree of pages belonging to clusters (IPACT). Lucas et al. (2012) suggested a new recommender system based on fuzzy logic and associative classification. In this paper, a CBA fuzzy algorithm was used to classify users and to apply association rules. This method uses collaborative filtering and content-based methods; therefore, it is a hybrid model. This method first uses other users' behavioral data and then uses the groups' properties and collaborative filtering methods. On the other hand, since proposed method uses the previous behavior of the active user to identify its classified group, it is a content-based approach. One of the major achievements of the associative classification in this study was having low amounts of false positive suggestions which are suggested to users, but do not attract their attention.

3. Usage mining based personalization architecture

The overall personalization process based on usage mining can be divided into two components. The offline component includes storage of data in a transaction file and special usage mining tasks (Mobasher et al., 2000), which include user clusters extraction in the present study. After the usage mining task is accomplished, the online component implements datasets and user clusters to offer suggestions based on their recent activities (Unler & Murat, 2010). Fig. 1 shows the structure of the suggested method. The tasks involved in each component of the proposed method are also explained in details.

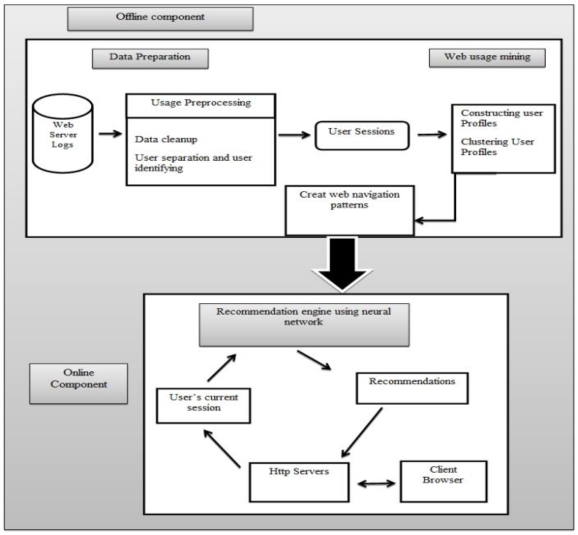

Fig. 1. The architecture of the proposed method

3.1.Pre-processing server records

In general, before applying web mining algorithms on records server, several pre-processing tasks have to be performed (Unler & Murat, 2010). For this research, these pre-processing tasks included data cleaning, separation and identification of user sessions.

3.1.1. Data cleaning

In web server records, not all of the registered records are suitable for web usage mining and unsuitable registries should be deleted (Tyagi et al., 2010). In this study, the following requests were removed:

- Requests sent by automated programs, such as crawling web node,
- Requests for image files that are associated with requests for specific pages,
- Registered records which correspond to undone requests,
- Registered records which include access methods except for “Get” and “Post”.

3.1.2.Identifying sessions

A user session is a collection of pages viewed by the user in the course of a visit to a website (Liu & Kešelj, 2007). Identifying user sessions from data records is difficult task, because it is possible that many users are using the same computer and one user might be using various computers. Therefore, the main problem is how to identify the user. In the case of websites, which require users to register, registering file includes users' system login information used to identify the user (Castellano et al., 2011). In our system, we have used IP addresses to identify user sessions. Each IP address is ascribed to a particular user. User session identification techniques are classified into two groups of time-based and content-based methods (Tyagi et al., 2010). In this method, after identifying users, the time based method was implemented for identifying sessions. In this method, some pages were considered sessions requested in a specific period of time (in our case 20 minutes).

3.2. Web usage mining

Following the pre-processing of web registries, web usage mining is executed for user sessions. Clustering, as an important tool in web usage mining, contributes to classification of users into clusters based on their mutual interests.

3.2.1. Session Vectorization

Let *p* be the collection of pages accessed by the user in web servers with $P = (p_j, j = 1, ..., m\}$ and each page has its own URL. Let *S* be the collection of user sessions with S= {s_i, I = 1, …, m} where each $s_i \in S$ is a subset of *P*. Each s_i session is shown by an m-dimensional vector like $s_i = \{w(p_1, s_i), (p_2, s_i), ..., (p_m, s_i)\}$ where $w(p_j, s_i)$ is identified for the j-th (1≤j≤m) viewed web page in session s_i. Note that web page $p_i \in P$ can be repeated in every $s_i \in S$ session. In order to weight the pages, it is necessary to identify user's interest in the page. Criteria adopted in the proposed system describe the amount of interest in each page accessed by the user as a function with three variables including view time, frequency of page access and date. The degree of user interest can be calculated by combining the above three criteria shown in Eq. (1):

$$W_{ij} = \frac{2 \times f_p \times t_p}{f_p + t_p} \times d_p \tag{1}$$

where, f_p represents page frequency, which means that in each session, it is possible that a user views a page for more than once and the more these views are, the more important that page is in the mentioned session compared with other pages. If N_{ij} is the number of user's accesses to page p_j and $\sum_{k=1}^{n_i} N_{ik}$ is total accesses, then:

$$f_p = \frac{N_{ij}}{\sum_{k=1}^{n_i} N_{ik}} \tag{2}$$

where t_p represents time, which refers to the time spent on a page. If users spend more time on a special page, that page is more favorite and if a page is not of interest to users, they will reject it and move on to other pages. We also need to consider this fact that quick movement to another page might be due to the small length of the page and this should be considered in the calculation of the page importance. Therefore, we have to change the properties of the time of page length or page bytes to normal. IF t_{ij} is the time spent by user on p_j page and size (p_j) is the size of p_j, then:

$$t_p = \frac{\frac{t_{ij}}{Size(p_j)}}{Max_{p_j \in P}(\frac{tij}{Size(p_j)})} \tag{3}$$

where $d_{p:}$ represents the Date, which is important because the possibility of requesting new pages by users is more than that of old pages and old pages are viewed less, because users are looking for new information. Hence, we considered page dates in calculating page weights and assumed that the more the page view date is close to present time, the higher would be the page weight. Moreover, if the page is older, it will get less weight. We have written Eq. (4) for these functions. If D_c is the present date, D_l is the date on which page is viewed by the user and $dif_d = d_c - d_l$, then:

$$d_p = \begin{cases} 1 & dif_d < a \\ 1 - (\frac{dif_d}{b}\ ) & b < dif_d < a \\ 0 & dif_d > b \end{cases} \tag{4}$$

3.2.2. Creating User Profiles

This system module is used to create user profiles. For this purpose, we have classified session vectors associated with different users obtained through session vectorization. We suppose that $\mathrm{s}_1, \mathrm{s}_2, \dots, \mathrm{s}_\mathrm{k}$ are the collection of sessions related to ith user ($\mathrm{u_i}$). Average vector s_{ui} for u_i user is calculated. In fact, this average vector is a representation of the user's favorite pages. The weight of each web page in average vector is calculated based on the average weight of that page in all user sessions ($\mathrm{s}_1, \mathrm{s}_2, \dots, \mathrm{s}_\mathrm{k}$).

3.2.3. Clustering User Profiles

In the proposed system, we used the clustering algorithms Fuzzy *C Means* and *S(c)* criteria (as shown in Fig. 2). S(c) criteria are meant to minimize spaces between data inside clusters and to maximize spaces between clusters which is described as follows (Tikk & Biró, 2001):

$$S(C) = \sum_{i=1}^{N}\sum_{k=1}^{C}(\mu_{ki})^m(||\mathrm{s}_{\mathrm{u_i}} - v_k||^2 - ||v_k - \bar{\mathrm{s}}_{\mathrm{u}}||^2) \tag{5}$$

where, N is the number of user profiles; C is the number of clusters, $C>2$; s_{ui} is the ith user profile; $\bar{\mathrm{s}}_{\mathrm{u}}$ is the average user profiles associated with kth cluster; v_k is the center of kth cluster (vector); μ_{ki} is the registering rate of ith input to kth cluster and m is the fuzzy exponent and m>1.

The process of algorithm is as follows:

The clustering algorithm:

```
Input: user profiles( $\boldsymbol{s}_{\boldsymbol{u}_i}$), Maximum Number of Cluster
Output : $\boldsymbol{c_j}(\mathbf{1} \le \boldsymbol{j} \le \boldsymbol{k})$
Step1: Set Number of Cluster C=2
Step2: Make Fuzzy C Means Clustering
Step3:Calculate S(C) criterion
Step4: set C=C+1
Step5:If C is Maximum Number of Cluster then go to Step6 Else go to Step2
Step6: If S(C) is Minimum for C then C is Optimal number of clusters
```

Fig. 2. FCM clustering algorithm with S(C) criteria

The result of profiles clustering is $C = \{c_1, c_2, \dots, c_k\}$ where for every $\mathrm{c_j}(1 \le \mathrm{j} \le \mathrm{k})$. we have the subset of user profiles in which k is the number of clusters. Every average vector shows the browsing pattern of users in a cluster in a special class of accessed web pages. Eq. (6) is used as a result of profiles clustering to show the total browsing patterns of users.

$$\mathrm{NP} = \{\mathrm{np}_1, \mathrm{np}_2, \dots, \mathrm{np}_\mathrm{k}\} \tag{6}$$

where, each np_i is a subset of P web pages. The vector of user browsing patterns shows a condensed view of the behavior of a group of users based on their common interests and information needs (Wu et al., 2013). These movement patterns are used to determine the similarity between the new profiles and previous ones.

4. Building a web recommender system using neural networks

The aim of this section is to collect user's current session and provide necessary suggestions to the active user. In this section we used a neural network to find the most similar clusters in user's current session and recommend appropriate pages. For this purpose, first we have to train the neural network. Navigational patterns obtained from previous stages are data sets used for training the neural network. The navigational pattern extracted from the previous components is considered as a neural network input. For this purpose, pages in navigational patterns are given to the neural network as input while network output is the number of clusters previously chosen for each navigation pattern. After training the neural network, we have to determine which active user belongs to which cluster. For this purpose, first we need to prepare the current user session in a suitable way for entering the neural network. Therefore, it is necessary to create user current session as profile mentioned in 3-2-1 section. Then in order to determine suitable cluster for the current session, we need to include the current session profile into neural network input. After detecting the suitable cluster number, those cluster pages which are not viewed in current session have a higher potential of being viewed as the next page by user. In this article, we also considered the effect of page rates in search engines in providing a suggestion list to the user. For this purpose, to those pages, which had better ratings in search engines, higher ranks were given. The suggestion list was analyzed by search engines based on the users' page interest rate and pages ratings and suitable pages were suggested to the user.

5. Case study

In this paper, we used data on the CTI -Depaul website. The data set included information on user sessions stored on CTI Depaul in 2002 for a two-week period (Barrueco Cruz & Krichel, 2002). In the proposed method, the following assumptions were made:

- Duration of a session is 20 minutes.

- Page size is equal to the number of bytes.

- The number of pages is 350.

The aforementioned clustering algorithm for grouping vectors into clusters based on user's behavior was also used. According to Fig. 3, the S (c) criterion for 6 clusters has the minimum value.

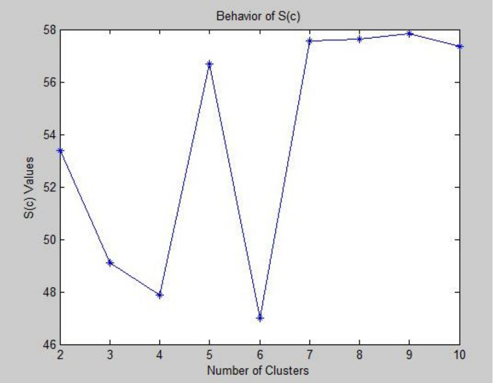

Fig. 3. Clustering User Profiles Diagram

After clustering, 750 data items were used to train the neural network. The remaining 250 items were used to test the system.

5.1. Assessment of the Proposed Method

In order to evaluate the proposed system, the precision and recall criteria were used. To this end, the following procedure was adopted:

Precision is the ability of the recommender system to generate precise suggestions. In other words, precision of suggest is the ratio of accurate suggestions to total suggestions (AlMurtadha et al., 2010).

$$\mathrm{precision}(\mathrm{rs},\mathrm{rp}) = \frac{|rs \cap rp|}{|rs|} \tag{7}$$

Recalling is the ability of the recommender system to generate all of the suggestions seen by the user (AlMurtadha et al., 2010).

$$recall(rs, rp) = \frac{|rs \cap rp|}{|rp|} \tag{8}$$

Fig. (4) and Fig. (5) show the average precision and recalling for the suggestions presented to several users using the proposed method. These figures show the result of the comparison between proposed method despite using date in user profile and the proposed method (without inclusion of date), the IPACT system proposed in (AlMurtadha et al., 2011) and user natural behavior.

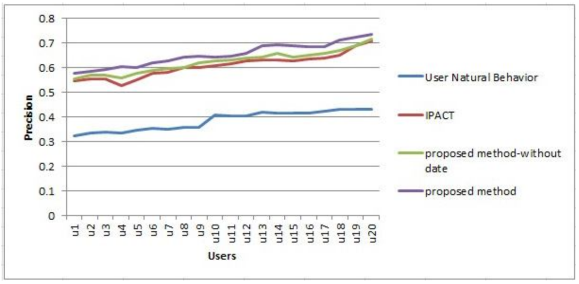

Fig. 4. Comparison Chart of the number of users using the precision criterion

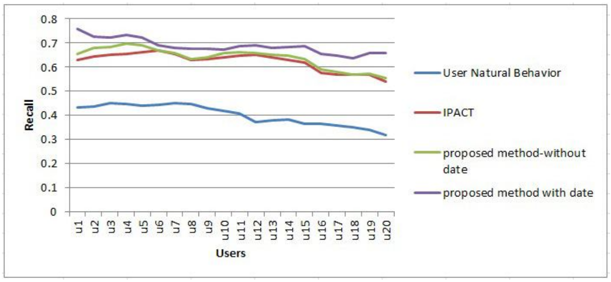

Fig. 5. Comparison Chart of the number of users using the recall criterion

6. Conclusion

This paper proposed a method to construct user profiles and to generate recommendations for user future requests. This paper used data from the user profile based on the frequency of access to the records server pages, time spent by the user on the pages and date of page views. We assumed that users are more likely to ask for newer pages as old pages are less favorable due to the interest of users in new information. For this purpose, we applied the page date factor for page weights, in a way that the more the date of visiting the page was close to the current date, the desired page received more weight. In addition, if the page was older, it was assigned a lower weight. After creating user profiles, using Fuzzy C-means clustering algorithm and the S (c) criterion, users with similar interests were classified. Finally, a neural network model was proposed to explore the proposed model and online suggestions were created for the active user using the suggestion module. Suggestions were evaluated based on the user's favored pages and page ranks in the search engines. Therefore, appropriate pages for active users were proposed. Research showed that the proposed method provides satisfactory precision in predicting user future requests.

References

AlMurtadha, Y. M., Sulaiman, M. N., Mustapha, N., & Udzir, N. I. (2010). Mining web navigation profiles for recommendation system. *Information Technology Journal*, *9*(4), 790-796.

AlMurtadha, Y., Sulaiman, M. N., Mustapha, N., & Udzir, N. I. (2011). IPACT: Improved web page recommendation system using profile aggregation based on clustering of transactions. *American Journal of Applied Sciences*, *8*(3), 277-283

Anand, S. S., & Mobasher, B. (2003, August). Intelligent techniques for web personalization. In *Proceedings of the 2003 international conference on Intelligent Techniques for Web Personalization* (pp. 1-36). Springer-Verlag.

Barrueco Cruz, J. M., & Krichel, T. (2002). Automatic extraction of citation data in a distributed digital library.. Proceedings of the 2nd International Workshop on New Developments in Digital Libraries, pp. 23-31.

Castellano, G., Fanelli, A. M., & Torsello, M. A. (2011). NEWER: A system for NEuro-fuzzy WEb Recommendation. *Applied Soft Computing*, *11*(1), 793-806.

Göksedef, M., & Gündüz-Öğüdücü, Ş. (2010). Combination of Web page recommender systems. *Expert Systems with Applications*, *37*(4), 2911-2922.

Janssens, F., Zhang, L., Moor, B. D., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, *45*(6), 683-702.

Liu, H., & Kešelj, V. (2007). Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. *Data & Knowledge Engineering*, *61*(2), 304-330.

Lucas, J. P., Laurent, A., Moreno, M. N., & Teisseire, M. (2012). A fuzzy associative classification approach for recommender systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *20*(04), 579-617.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on Web usage mining. *Communications of the ACM*, *43*(8), 142-151.

Mustapaşa, O., Karahoca, D., Karahoca, A., Yücel, A., & Uzunboylu, H. (2010). Implementation of semantic web mining on e-learning. *Procedia-Social and Behavioral Sciences*, *2*(2), 5820-5823.

Nasraoui, O., Soliman, M., Saka, E., Badia, A., & Germain, R. (2008). A web usage mining framework for mining evolving user profiles in dynamic web sites.*Knowledge and Data Engineering, IEEE Transactions on*, *20*(2), 202-215.

Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C. D. (2003). Web usage mining as a tool for personalization: A survey. *User modeling and user-adapted interaction*, *13*(4), 311-372.

Taghipour, N., & Kardan, A. (2008, March). A hybrid web recommender system based on q-learning. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1164-1168). ACM.

Tyagi, N. K., Solanki, A. K., & Wadhwa, M. (2010). Analysis of Server Log by Web Usage Mining for Website Improvement. *International Journal of Computer Science Issues*, *7*(4), 17-21.

Tikk, D., & Biró, G. (2001). Sugeno-Yasukawa fuzzy modelling: survey and improvements. 2nd *International Symposium of Hungarian Researchers on Computational Intelligence, Budapest*, 175-186.

Unler, A., & Murat, A. (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research*, *206*(3), 528-539.

Wu, J., & Wu, Z. (2013). Improved fuzzy c-means clustering for personalized product recommendation. *Research Journal of Applied Sciences, Engineering and Technology*, 6(3), 393-399.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, *16*(3), 645-678.