Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

Prediction of users' future requests using neural network

Seyedeh Foroozan Rashidi[a*], Ali Harounabadi[b] and Mashaalah Abasi Dezfouli[a]

[a]*Department of computer engineering, Science and Research Branch, Islamic Azad University, khouzestan-Iran*
[b]*Department of computer engineering, Islamic Azad University, Central Tehran Branch, Iran*

ARTICLEINFO

Article history:
Received March 25, 2012
Received in Revised form
April, 30, 2012
Accepted 30 May 2012
Available online
June 7 2012

Keywords:
Web usage mining
Neural Network
Clustering
Recommender engine

ABSTRACT

With the rapid growth of the World Wide Web, finding useful information from the Internet has become a critical issue. Automatic classification of user navigation patterns provides a useful tool to solve these problems. In this paper, we propose an approach for classification of users' navigation patterns and prediction of users' future requests. Users' profiles are constructed based on Web log server files and one of clustering methods is implemented to users' profiles for assigning navigation patterns. Finally, using neural network, recommender engine produces a relevant recommendation list of web pages to the active user. The preliminary results indicate that the proposed approach has high accuracy and coverage in prediction of users' future requests.

1. Introduction

With the explosive growth of knowledge available on the World Wide Web, which lacks an integrated structure or schema, it becomes much more difficult for users to access relevant information, efficiently. Meanwhile, the substantial increase in the number of websites presents a challenging task for webmasters to organize the contents of the websites to cater to the needs of users (Mobasher, 2004; Liu & Keselj, 2007; Albadvi & Shahbazi, 2009). These problems have made Web personalization an indispensable tool for both Web-based organizations and for the end users. Web personalization can be described as any action, which makes the Web experience of a user customized to the user's taste or preferences (Mobasher, 2004). Recently, Web mining techniques have been widely applied for personalization (Guandong, 2008). The classification of navigation patterns can enhance the quality of personalized web recommendations, which aims to predict which web pages are more likely to be accessed next by current users. The basis of our approach is to extract navigation profiles, which capture similar behavior of site users. The navigation pattern can be used for predicting the navigation behavior of current users, thus aiding in web personalization. Recently, web usage mining techniques have been widely applied for discovering interesting and

* Corresponding author. Tel: +989378637844
E-mail addresses: foroozan.rashidi@gmail.com (F. Rashidi)

doi: 10.5267/j.msl.2012.06.007

frequent users' navigation patterns from Web server logs. Sequential pattern mining (Zhou et al., 2004), association rule mining (Lin et al., 2000; Mobasher et al., 2001) and clustering (Mobasher et al., 2001), discover different access patterns from web logs, which can be modeled and used to offer a personalized and proactive view of the web services to users.

In this paper, we propose an experimental system, which uses web usage mining and could result in a more accurate classification of user navigation patterns, and consequently lead to a more accurate prediction of users' future requests. We build users' profile according to obtained information from web server logs and during building user profiles we consider users' behavior history. Then we build users' navigation pattern to predict future users' requests. Then, using neural network, recommender engine produces a relevant recommendation list of web pages to the active user. Results of the implementation indicate that the proposed approach can predict and categorize the users' navigation behavior with high accuracy and coverage. The rest of this paper is organized as follows: In Section 2, we review recent research and section 3 describes the research background; Section 4 describes the architecture of the proposed system. Section 5 presents experimental results assessing the performance of our system. Finally, Section 5 concludes this paper with suggestions for future research.

2. Related works

Web mining is the mining of data related to the World Wide Web categorized into three active research areas based on the mining goals: *Web content mining*, *Web structure mining*, and *Web usage mining* (Dunham, 2003; Srivastava et al., 2000; Madria et al., 1999). It includes content mining, which is the process of extracting knowledge from the content of websites and structure mining, which uses links and references within web pages. Usage mining, also known as web-log mining, which studies user access information from logged server data in order to extract interesting usage patterns (Dunham, 2003).

Recommender systems generally consist of two phases: offline pattern extraction and online recommendation (Li & Zaiane, 2004). In offline recommendation phase, the web usage mining techniques are applied to reveal the hidden navigation patterns of users that stored in the web server logs. In the online phase, the current session of active user is compared with these navigational patterns with some similarity measures and consequently recommends pages are determined (Mobasher, 2004*)*. Most research activities in web mining have centered on content mining and usage mining (Liu & Keselj, 2007*)*. A project aiming at extracting navigation behavior models of a site's visitors was introduced in (Baglioni et al., 2003*)* but its classification accuracy was not promissing. Nakagawa and Mobasher (2003) proposed a recommender system, which adopts a clustering technique to obtain both the site usage and site content profiles. In this work, the authors use association rule mining and sequential pattern mining to generate navigational patterns of Web users. A switching (Burke, 2002*)* hybridization method was used to integrate the navigational patterns of Web users in order to generate a recommendation set. Goksedef and Gunduz-Oguducu (2007) proposed a recommendation model called consensus recommender where several recommendation models based on Web usage mining techniques are integrated. They showed that consensus model achieves a better prediction accuracy compared to its individual components.

3. Backgrounds

In the development of proposed model of this paper, some technologies are required such as the clustering and the neural network. These will be discussed in the following sub-sections.

3.1. Clustering

There are large majority of methods used for pattern discovery from Web data, which are based on clustering methods. Clustering aims to divide a data set into groups, which are very different from each other and whose members are very similar to each other. The k-means is the simplest and most commonly used algorithm. A general k-means has the following steps: (1) initialize the clusters' centers by seed selection; (2) group each data point into the nearest cluster center based on distance calculation; (3) recalculate each center using the mean of all the points in the same cluster; (4) move some data points from one cluster to another to minimize the sum-of-squares criterion, back to (2) until convergence condition is met (He et al., 2003).

3.2. Neural Network

In general, artificial neural networks are composed of interconnecting artificial neurons (nodes). One of the most popular and most powerful architectures is the Multilayer perceptron. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. MLP utilizes the backpropagation for training the network. Learning occurs in the Multilayer perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result (Basheer & Hajmeer, 2000)

4. System design

In this work, we have designed an experimental system to predict future requests of users using web usage mining. Our system consists of five major modules. In the following subsections, we will describe the algorithms for each component of the system in detail.

4.1 Web Log Preprocessing

Web log preprocessing aims to reformat the original web logs to identify all web access sessions. Generally, several preprocessing tasks need to be done before performing web mining algorithms on the web server logs. For our work, these include data cleaning, user differentiation and session identification (Cooley et al., 1999). The identified sessions are further split up into two sets based on the date of log entries: “training set” and “testing set”. The training set is used to build user profiles, while the testing set is prepared for evaluation and experiment of system.

4.2 Session Vectorization

We can represent a user session as a collection of transactions, which includes a series of weighted pages, during the visiting period. From such viewing point, we generate the following user session expression. Let P be a set of web pages accessed by users in web server logs, $P = \{p_1, p_2, ..., p_m\}$, each of which is uniquely represented by its associated URL. Let S be a set of users access sessions. Hence, $S = \{s_1, s_2, ..., s_n\}$, where each $s_i Є S$ is a subset of P. We represent each session s_i as an m-dimensional vector over the space of web pages, $s_i = \{w(p_1, s_i), w(p_2, s_i), . . . , w(p_m, s_i)\}$, where $w(p_j, s_i)$ is a weight assigned to the jth web page $(1 \leq j \leq m)$ visited in the session s_i. In this paper, the weight $w(p_i, s_j)$ is defined as the interest degree of a particular user to the page, which is the harmonic mean of Frequency and Duration to represent this interest (Dumais, 2003*)*, and shown as Eq. (1):

$$Interest(Page) = \frac{2 \times Frequency(Page) \times Duration(Page)}{Frequency(Page) + (Duration(Page)} \tag{1}$$

At the end, every user access session is transformed into an m-dimensional vector of weights of web pages, i.e., $s = \{w_1, w_2, ..., w_m\}$. For reducing dimensions, we here use a frequency threshold f_{min} to filter out web pages that are accessed less than f_{min} times.

4.3 Constructing Users Profile

This system module attempts to construct users' profiles to distinguish different users session vectors which are obtained from Session Vectorization. Let $\{s_1, s_2, ..., s_k\}$ be a set of session vectors of i^{th} user (u_i). We compute a mean vector s_{ui} for the user u_i as its representation. This mean vector represents web pages, which are interesting in by the users. The weight of each web page in the mean vector is computed by the average weight of the web pages across total access sessions of the user {s_1, s_2, ..., s_k}. During the mean vectors calculation, we consider users behavior history. Therefore, in the profile constructing module, a linear incremental weight is given to the accessed web pages according to the access sequence of the web pages.

4.4 Clustering Profiles

Given the transformation of users profile into a multi-dimensional space as vectors of web pages, standard clustering algorithms can partition this space into groups of profiles that are close to each other based on a distance measure. In this paper, the well-known *K-means* algorithm is applied to cluster users' profiles. Profiles clustering results in a set of clusters, $C=\{c_1, c_2, ..., c_k\}$, which each $c_i(1 \leq i \leq k)$ is a subset of the set users' profiles, and k is the number of clusters. We compute a mean vector for each cluster $c \in C$ as its representation. We use a weight threshold, w_{min}, for the mean vector of each cluster. Web pages remained in each cluster are considered of more interest to users. Each mean vector represents the representative users' navigation patterns of a cluster in which a particular set of web pages are accessed. As the results of profiles clustering, $NP = \{np_1, np_2, ...,np_k\}$ is used to represent the set of users navigation patterns, in which each np_i is a subset of P, the set of web pages.

4.5 Recommender engine using neural network

The task of this component is receiving the user's current access session and producing a recommendation list for the active user. We use Neural Network to find the most similar cluster to the user's current access session and recommend appropriate pages to the user. Therefore, we train Neural Network using the navigation patterns. The navigation patterns have been considered as the inputs of the network and the relevant cluster's number as the output of the one. The network may be trained with these data. We construct a profile for the users' current session. The profile session is a vector of weights of web pages visited in the session. We need to determine the cluster to which the session profile most likely belongs. In order to, input Layer receives the session profile and the network determine relevant cluster's number for the session. Therefore, web pages in the cluster that have not been accessed have great potential to be the next pages that the user wants to see, therefore they will include in recommendation list.

5. Experimental Setup

In our experiments we used a publicly available data set. This data set includes the sessionized data for the DePaul University CTI web server based on a random sample of users visiting the site for a two week period during April 2002 (Maya, 2002). The data set includes 213 distinct page and 13745 distinct user sessions of length more than one. We split the sessions in two non-overlapping time windows to form a training (9745 sessions) and a test (4000 sessions) data set.

5.1. Evaluation

Our evaluation methodology is as follows. Each test session *ts* in the test session set is divided into two parts. The first *n* web pages of session in *ts* are used for generating recommendations, and the second part is simulated as the future requests (page visits) which are compared with the output of the recommendation system. The value *n* reflects the maximum allowable window size for the experiments (in our case w=4) (Mobasher et al., 2001). This size represents the last *w* pages in the first part of session called active session window (*asw*). For testing session *ts* with size n≤4, active session window *asw* with size n-1 were chosen. The recommendation engine takes *asw* and the recommendation threshold μ as the input and generates a recommend list which denoted by $R(asw, \mu)$. Note that $R(asw, \mu)$ contains all pages whose recommendation score is at least μ. The set of pages $R(asw, \mu)$ can now be compared with the remaining $|ts|-n$, pages in *t*. We denote this portion of *ts* by $eval_{ts}$. Our comparison of these sets is based on 2 different metrics, namely, precision and coverage (Mobasher et al., 2001), which are defined as Eq. (2) and Eq. (3):

$$precision\big(R(asw,\mu)\big) = \frac{|R(asw,\mu) \cap eval_{ts}|}{|R(asw,\mu)|} \quad (2)$$

$$coverage(R(asw,\mu)) = \frac{|R(asw,\mu) \cap eval_{ts}|}{|eval_{ts}|} \quad (3)$$

Precision measures the degree to which the recommendation engine produces accurate recommendations. Coverage measures the ability of the recommendation engine to produce all of the pages that are likely to be visited by the user.

5.2. Experimental Results

Weka machine learning toolkit is used to perform the K-means algorithm and the Euclidean distance is adopted as the distance measure used for clustering (Bouckaert, et al. 2008). The optimal number of cluster compactness and cluster separation proposed in (He et al., 2003). The clustering method applied and produced 22 clusters. A perceptron network was used to learn from the data. The development of an ANN requires partitioning of the parent database into three subsets: training, test, and validation. We used 60% of all data (training data set) for training, 10% for validation and remaining data for testing the network. We might use one or two hidden layer in the network and then trained it with various neurons in each layer in MATLAB software environment.

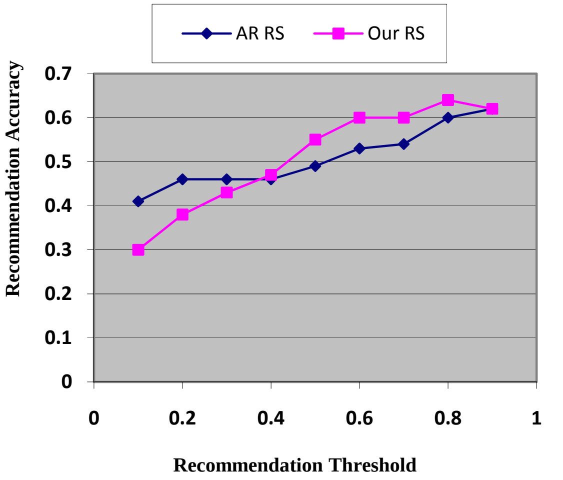

Fig 1. Recommendation accuracy comparison: Our system vs. association rule-based system

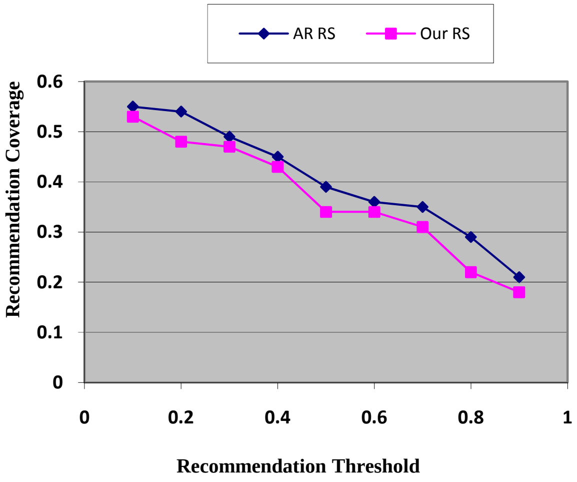

Fig. 2. Recommendation coverage comparison: our system vs. association rule-based system

Fig. 1 and Fig. 2 show the accuracy and coverage of the system. Furthermore, association rule based recommender systems were also implemented with the same data sets. The results were also depicted in Fig. 1 and Fig. 2 for comparison. Fig. 1 shows better results for recommendation accuracy of association rule based recommendation system for threshold values of 0.1, 0.2 and 0.3, but better results for our system in other thresholds (0.4 to 0.9); and Fig. 2 indicates our recommendation system results are close to association rule based recommendation system. According to (Mobasher, 2004), it is notable that the association rule mining based recommendation system has the best performance in coverage metric among the web usage mining based CF recommendation systems.

5. Conclusion

In this paper, we have proposed an approach to prediction of users' future requests by using an artificial neural network. We first built users' profiles according to obtained information from web server logs. The proposed model of this paper then built users' navigation pattern for prediction of future users' requests. Finally, using neural network, recommender engine produced a relevant recommendation list of web pages to the active user. The results of the implementation indicate that the proposed approach has high accuracy and coverage in prediction of users' future requests.

References

Albadvi A. & Shahbazi M. (2009). A hybrid recommendation technique based on product category attributes, Expert Systems with Applications, 36(9), 11480-11488.

Baglioni M., Ferrara U., Romei A., Ruggieri S. & Turini F. (2003). Preprocessing and mining Web log data for web personalization, In: AI*IA, 237–249.

Basheer, I. A., & Hajmeer, M. (2000), Artificial neural networks: Fundamentals, computing, design and application. *Journal of Microbiological Methods*, 43, 3–31.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331–370.

Cooley, R., Mobasher, R. & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1), 5–32.

Dumais, S., Joachims, T., Bharat, K. & Weigend, A. (2003). Workshop Report: Implicit Measures of User Interests and Preferences, ACM SIGIR Forum.

Dunham, M. H. (2003). *Data Mining: Introductory and Advanced Topics*. Prentice Hall.

Goksedef, M. & Gunduz-Oguducu, S. (2007). A Consensus Recommender for Web Users, In: ADMA. The 3rd international conference on advance data mining and applications, 287–299.

Xu, G. (2008). Web Mining Techniques for Recommendation and Personalization, PhD thesis, Victoria University.

Jain, A.K., Murty, M.N., & Flynn, P.J. (2000). Data clustering: A review. *ACM Computer*, 31(3) 264-323.

Li, J. & Zaiane, O. R. (2004). Combining usage, content, and structure data to improve web site recommendation. In: Proceedings of 5th international conference on electronic commerce and web, 305–315).

Lin, W., Alvarez, S. A., & Ruiz, C. (2000). Collaborative recommendation via adaptive association rule mining. *Web Mining for E-Commerce – Challenges and Opportunities, Second International Workshop, Boston*, MA, USA.

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

Madria S. K., et al., (1999), Research Issues in Web Data Mining. In: Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99. 1999, p. 303-312, Florence, Italy.

Mobasher, B., Dai, H., Luo, T. & Nakagawa, M. (2001). Effective personalization based on association rule discovery from web usage data. *In: Proceedings of the 3rd International Workshop on Web Information and Data Management, ACM Press, Atlanta, GA, USA*, 9–15.

Mobasher B. (2004). *Web Usage Mining and Personalization, Practical Handbook of Internet Computing*. Chapman Hall and CRC Press.

Nakagawa, M. & Mobasher, B. (2003). A hybrid web personalization model based on site connectivity. Proceedings of WebKDD, 59–70.

Phatak, D. S., & Mulvaney, R. (2002). Clustering For personalized mobile web usage. *Proceedings of the IEEE FUZZ'02, Hawaii, USA*, 705–710.

Srivastava, J. Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2), 12-23.

Zhou, B., Hui, S.C. & Chang, K. (2004). An intelligent recommender system using sequential Web access patterns. *Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems, Singapore*, 1–3.

Bouckaert, R.R., et al. (2008), WEKA Manual for Version 3-6-0.

He, J., Tan, A. H., Tan, C. L., & Sung, S. Y., (2003). On quantitative evaluation of clustering systems. *Information Retrieval and Clustering*, 105–134.