ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

A new algorithm to create a profile for users of web site benefiting from web usage mining

Masomeh khabazfazli[1,2], Dr. Ali Harounabadi[3] and Dr. Shahram Jamali[4]

[1] Department of computer, Ardabil Science and Research branch, Islamic Azad University, Ardabil, Iran
[2] Department of computer, Ardabil Branch, Islamic Azad University, Ardabil, Iran
Ma.fazli64@gmail.com
[3] Department of Computer Science, Islamic Azad University, Central Tehran Branch, Tehran, Iran
A.harounabadi@gmail.com
[4] Computer Engineering Department, University of Mohaghegh Ardabil, Ardabil, Iran
Jamali@iust.ac.ir

Abstract

Upon integration of internet and its various applications and increase of internet pages, access to information in search engines becomes difficult. To solve this problem, web page recommendation systems are used. In this paper, recommender engine are improved and web usage mining methods are used for this purpose. In recommendation system, clustering was used for classification of users' behavior. In fact, we implemented usage mining operation on the data related to each user for making its movement pattern. Then, web pages were recommended using neural network and markov model. So, performance of recommendation engine was improved using user's movement patterns and clustering and neural network and Markov model, and obtained better results than other methods. To predict the data recovery quality on web, two factors including accuracy and coverage were used

Keywords: *Web Page Recommendation, Web Mining, Web Usage Mining, Clustering, Neural Network, markov model*

1. Introduction

Upon establishment and expansion of internet and subsequently websites enhancement and upraise of its applications by users, searching the contents among extensive information on internet pages has become difficult for web users. The users face a great volume of recovered data. On the other side, topics such as purposeful advertisements and awareness of users' information are very important. Therefore, web mining is propounded for solving this problem. Web mining is the process of unknown and useful knowledge discovery through web data. Currently, broad researches have been applied in this relation and their purpose is solving problems related to data recovery [1].

One of objectives of this study is web pages recommendation for users and time and cost saving as well as better support of purposeful advertisements and electronic business. Thus, upon using web mining methods, this group of problems are solved somewhat. Users may select pages recommended to them which are related to the subject. In this part, generalities, objectives and necessities of this research w analyzed. In second part, background of study is explained. Third part includes main idea for offering web recommendation engine based on web usage mining. In fourth part, evaluation and results of experiments are provided and compared to other methods, and in final part summary of paper is provided.

2. RELATED WORK

Recently, web usage mining techniques are used extensively for prediction of internet pages. Access patterns are discovered from record file using methods such as association rules mining, clustering etc. that is used for prediction of users' behavior. In [2], a web usage mining method was used offered therein clustering was used so that users' behavior was clustered based on measurement set of log file data similarity to be used for prediction of internet pages. In fact, clustering has been made using similarity of above approximation. In this process, clustering was provided based on subject and shows common interests in each cluster. In [3], web usage mining techniques were used and analyzed the problems from two aspects including improvement of search engine through static saving of search results and weblog posts. This study offered search coverage method and used graph for recommendation to users. In [4], rough set theory was used for log file processing and keywords, upon combining two methods of content and collaborative filtering based recommender systems through design of two-layer graph that was made along with graph partitioning. Each node in pages later and users' layer respectively shows web pages and users. Therefore, similarity between pages and users is obtained by this way in graph partitioning.

3. Background

In this part, background contents which are important for understanding the offered method are raised. At first, web mining, then personalization based on usage mining and at end clustering approaches and neural network and markov model will be explained.

3.1 Web mining

Web mining is a subset of data mining technique for covering the web patterns that based on web pages analysis and which part of web data is explored is divided in three parts including web structure mining, web content mining and web usage mining [5].
Web content mining is discovery and extraction of useful data from web pages. Web structure mining discovers and analyzes the model. In this kind of web mining, web is modeled as a graph therein web pages are assumed as graph nodes and links between pages as graph edges [5]. Web usage mining includes discovery of user's access patterns for web server recording file.

3.2 Personalization based on web usage mining

Personalization is as one of the application fields of web exploring by which pages contents can be changes in accordance with users' interests in order to provide the internet services in a better way and also to meet the needs of users quickly. Several web personalization systems have been created based on web mining that all of them include two main stages: in the first stage which is performed offline, training data taken from user behavior on the Web are explored in order to detect the access patterns and extract the users' model. In the second stage which is done online, the model extracted from the first stage is used for interpretation and comparison with the traversal pattern of active user and then propositions are provided based on this comparison. The purpose of web personalization based on exploring the web application is offering a set of objects to the current user with orientation towards the user's preferences and interests.

3.3 Clustering

Clustering or cluster analysis is the process of grouping some physical and virtual objects in classes of similar objects. A cluster includes a set of data objects which are similar to each other. In general, two types of clustering (transaction clustering and page visiting clustering) can be applied for the transaction data of web application.

Each of these approaches has different applications and in particular, both of the two approaches can be used for web personalization. K-Means algorithm is one of the most important clustering algorithms that are widely used. In this algorithm, the samples are divided into k clusters and the number of k has already been specified.

3.4 Neural network

Neural networks are available for simulating the human brain performance in remembering the information and learning. Human brain consists of a great number of nerves. Each of these nerves is connected to the other ones and sends signals to each other. Although each neuron has no the complex structure, but the set of these neurons create a more complex network. In fact, the artificial neural networks are going to create a special output by the special input and according to this, the concept of training or adjustment and learning of artificial network is achieved.

3.5 Markov model

Viewing Web transactions in the form of a series of page views allows that a number of useful models can be used to detect and analyze the user circulation patterns. One of the events is modeling the behavior of the user's circulation by Markov chain in the website. A Markov model with a set of states $\{S_1, S_2, ..., S_n\}$ and a transfer matrix is shown [6]:

$\{P_{1,1},\ldots,P_{1,n},\ldots,P_{2,1},\ldots,P_{n,1},\ldots,P_{n,n}\}$
Where $P_{i,j}$ is the probability of transition from state S_i to state S_j .

4. THE PROPOSED METHOD

In the proposed system, firstly data recorded in server weblog which are as the input data of system are preprocessed. Then, the web usage mining operations is performed on the identified sessions for building the functional patterns of user. Finally, web pages related to users' traversal method in the site are proposed based on their functional patterns. The web page is proposed by the neural Network and markov model. Following of this section will studied each of above components.

4.1 Log file preprocessing

There are various data with different formats in high volumes on the level of web. These data include HTML pages, CSS & JS files and variety of multimedia images [7]. In web usage mining, the preprocessing includes identifying users and their sessions that are used as main elements to detect the pattern [5]. An accurate identification of users and their sessions has a particular importance in web personalization because the users' models are made based on their behavior which they also themselves will be available as users' sessions.

4.2 constructing the users' session vectors

Now that we have preprocessed recorded data in the web server logs and have also achieved the proper data in the form of users' sessions, we are ready to accomplish the web usage mining.
Suppose that P equals to the set of pages accessed by users of a site according to the following definition:
$P=\{P_1,P_2,\ldots,P_m\}$
And each of the pages P_i $(1\le i\le m)$ have a specific and unique URL and S which represents the users sessions is expressed as following:
$S= \{S_1, S_2,\ldots,S_m\}$, $S_i \subset P$ $(1\le i\le m)$
Every Si session is demonstrated as an m-dimension vector:
$S_i=\{w(p_1, s_i), w(p_2, s_i), \ldots, w(p_m,s_i)\}$
In the above relation, the value of W (P_j, S_i) equals to the weight assigned to j-th visit in Si session and the value of j is among 1 to m. It should be noted that each of the above pages can be present in each of the sessions. In the above relation, the values of w must indicate the rate of users' interest in pages. One of the cases which have a relation with the user's interest in pages is the page frequency. The frequency of a page means the amount of access to that page in a session by users and it has a direct relation with the users' interest in that page. The following relationship shows the calculation of the rate of a page frequency in a session. In the following relationship, N-visited represents the number of a page visits in a session and visited pages equals to the whole set of the visited pages in a session. It is given by (1):

$$frequency(page)=\frac{N\_visited(page)}{\sum_{page\in visitedpages} N\_visited(page)} \quad (1)$$

One of the parameters that can specify the rate of user interest in a page, is the time duration spent by a user for visiting a page. The amount of time duration is a normalized value (between 1 and 0) which is calculated for each page by relation (2):

$$duration(page)=\frac{TotalDuration(page)/Length(page)}{Max_{page\in visitedpage}(TotalDuration(page)/Lengh(page))} \quad (2)$$

Another criteria which is effective on the rate of user interest in a session is the date of page visiting, this means that the pages which have been recently visited show a better reflection of user interests. But since the date of visiting is not a numerical value rather is a historical figure, some operations must be applied on it. The normalized numerical value of the date of page visiting is calculated by the following relation. In this relation, Date Origin, Date(Page) and PageLast are equal to the origin date, date of page visiting and last visited page in session, respectively. The amount of Date value is a normalized value (between 1 and 0) which is calculated from relation (3) for each page:

$$DateValue(page)=\frac{(Date(page)-DateOrigin)/(Date(pageLast)-DateOrigin)}{Max_{page\in visitedpages}((Date(page)-DateOrigin)/(Date(pageLast)-DateOrigin))} \quad (3)$$

Now, we use the harmonic mean to combine these three parameter. The value of Interest (page) that equals to the amount of user interest is obtained as (4). Let Da(page) be DateValue(page), F(page) be ferequency(page) and Dur(page) be Duration(page). So we have in (4) relation:

$$\text{int}erest(page)=\frac{3*F(page)*Dur(page)*Da(page)}{F(page)*Dur(page)+Dur(page)*Da(page)+F(page)*Da(page)} \quad (4)$$

The above value is a normalized value (between 1 and 0) that is calculated for each page.

4.3 creating the users profile

Here the user profile is an exhibitor of the resultant of his favorite pages. The users' profiles are created in this section. For doing this, first the set of sessions of all users

is separated by the separation of the user. Assume that Si ($1 \le i \le k$) is the set of the sessions of user U_i. The mean vector Sui is specified as the resultant vector for user U_i and will be indicative of the mean of user's interests in pages.
The weight of each web page in the mean vector is calculated by the average weight of that page in all the sessions of the user (S_1, S_2, ...,S_k). The user behavior history is also considered in calculating the mean vector of the session.

4.4 Clustering User Profiles and Neural Network the session

After obtaining the users profile which indicates the abstract of the users' interests in web pages, we can divide them into some groups using the clustering algorithm so that users' interests in web pages can be better organized and also it can be provided a background for extracting their internal patterns. For clustering the users' profile, we use k-mean algorithm. The result of the clustering is as k= {k_1, k_2,...,k_c} and c equals to the number of identified clusters by k-Means algorithm. Supposing that P is the set of users' profile, then we have the following formula per k_i Є k (k_i C P).

Since the weight of vectors is between 1 and 0, then their average value is also between 1 and 0. To reduce the number of dimensions, it is defined a threshold for pages weight in mean vectors of clusters. The pages whose weight is less than that threshold, are removed from the mean vector. And the remaining pages represent the most interests of users in the relevant cluster.
If we consider the set of users' movement patterns existing in website as a NP Set, then this set will be displayed as the following.
NP= {np_1, np_2,… , np_k}
In the above set, there is a mean vector for each cluster and we have for each member of the above set that each member of np_i is a subset of the set of website pages. These patterns are applied for determining the similarities between new profiles and the previous profiles.
Here, the neural network is used to find the closest cluster to the user session and to propose some proper pages to him. Network training is performed using the movement patterns obtained from the previous stages. Then, we should prepare the session of current user so that is appropriate for the neural network. Since the input of neural network is related to some weights of pages, we should create a profile for the current user based on pages weight. Now, we must determine that the new profile belongs to which one of the existing clusters. For this purpose, it is enough to give the new profile to the neural network until it can determine an appropriate number of clusters for this new profile.

4.5 Recommend pages using the Markov model in cluster

After using neural networks, was diagnosed nearest cluster to the user's current session, the next step is proposed from inside the cluster to new users, we are extracted sequences of pages that visited by Markov model. And through it propose the page with the highest Means page the most probable repeat to user. for do this we apply the markov model to training clustering data and suggest do to testing data to measure system performance, and use of the system to suggest to new users.

5. RESULT SIMULATION

In this thesis, it has been used the web server pages of Saskatchewan University for conducting the research. The data of this web server have been used as a set of web server logs which are derived from the two-week log data of the web server of the university site in 2004. This file includes 1480 user sessions and 570 pages which have been accessed by different website users during the mentioned sessions. A time period about 1800s was considered as the threshold time to extract the users sessions from the web server logs. Some pages which have been referenced less than 10 percent or more than 80 percent of the maximum frequency of access to pages were excluded according to [5]. During two stages, first 156 and then 46 website pages remained respectively, and thus 46 pages remained finally. Then, all the sessions with a length of less than 8 were removed, ultimately it remained about 617 sessions. After removing the inappropriate pages and sessions, we divided the obtained sessions into two parts. We used the first and second sessions as training sessions and test sessions, respectively. The first part is used for learning and the second part is applied for the system evaluation.
Clustering the profiles has been performed using the k-Means algorithm in the visual studio 2010 Software. Markov model has been performed in the visual studio 2010 software, and neural network has been performed in the matlab software. Thus, first the system was trained using the set of training data , then we used the test data set and created the new sessions for users by some test data sets without any role in creating the training profiles. Finally, we evaluated the efficiency of our system by these new data.

System is evaluated by two metrics including accuracy and coverage [8]. The value of accuracy is defined as a ratio of correct propositions to the whole propositions. In the other word, if the proposed equals to the set of proposed pages, R equals to the set of correct pages and size is a function

which implies the size of a set, then the value of accuracy is calculated by the (5) relation:

$$precision(proposed) = \frac{Size(precision \cap R)}{Size(R)} \quad (5)$$

Coverage is used for reviewing this issue that how many correct pages in the provided propositions have been covered by system. Its mathematical definition is as (6):

$$Coverage(proposed) = \frac{Size(proposed \cap R)}{Size(T)} \quad (6)$$

Given that the window size is usually three to four pages in papers and also we deleted the sessions with less length, then we considered the window size equal to 4 in this research.

In order to assess the proposed system based on the clustering by k-Means algorithm, we compared it with a method based on association rules and other method [9], their results have been shown in the two following figures.

Fig. 1 Compare the proposed method with other methods .

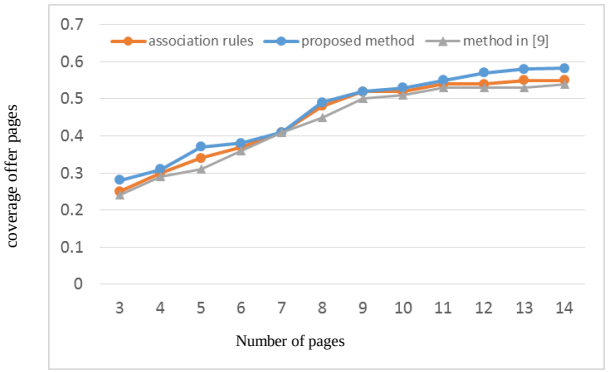

Fig. 2 Compare the proposed method with other methods.

It is completely clear that the value of accuracy has an inverse proportion to the number of proposed pages, so that the value of accuracy will reduce by increasing the number of proposed pages. This means that the number of proposed correct pages will be less than the whole pages proposed by the system.

As it is clear in the above figure, the level of coverage of both systems (proposed and based on K-Means) has a direct proportion to increasing the number of pages and also the level of coverage of both systems is increasing followed by raising the number of the proposed pages.

6. Conclusions

In this paper, a method was offered for prediction of web users' subsequent page selection using neural network and markov model. Our offered system benefits from web usage mining for recommendation to users. To achieve the users' survey pattern, a method has been used that firstly users' profile is made based on data extracted from web server records and during profile making process, history of users' behavior and page visiting date is taken into account. Then, upon clustering the profile, users' movement patterns are extracted. After obtaining movement patterns of recommendation engine using neural network and markov model, a list of suitable pages is recommended to the user. Summary of implementation shows that recommendation engine offered in this paper has appropriate accuracy and coverage for prediction of subsequent requests of user.

References

[1] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works, Journal of Expert Systems with Applications", Vol. 41, No. 4, Part 1, March 2014, pp.1432-1462.

[2] K., Santhisree, A., Damodaram, "Clustering on Web usage data using Approximations and Set Similarities, International Journal of Computer Applications", Vol. 1, No. 4, , 2010, pp.0975 – 8887.

[3] Ida Mele, "Web Usage Mining for Enhancing Search-Result Delivery and Helping Users to Find Interesting Web Content", ACM, WSDM'13, February 2013, pp.765-769.

[4] J., Jose, P., Sojan Lal, "Extracting Extended Web Logs to Identify the Origin of Visits and Search Keywords", Intelligent Informatics Advances in Intelligent Systems and Computing, Vol. 182, 2013, pp.435-441.

[5] G., Castellano, A. M., Fanelli, M. A., Torsello, "NEWER: A system for Neural-fuzzy Web Recommendation", journal of Applied Soft Computing, Elsevier Science Publishers B. V. Amsterdam, The Netherlands, vol. 11, No. 1, 2010, pp.793-806.

[6] X., Dongshan, S., Junyi, "A New Markov Model for Web Access Prediction", journal of Computing in Science and Engineering, vol. 4, no. 6, 2002, pp. 34-39.
[7] K., Goseva-Popstojanova, G., Anastasovski, A., Dimitrijevikj, R., Pantev, B., Miller, "Characterization and classification of malicious Web traffic, Journal of Computers & Security", Vol. 42 , May 2014, pp.92-115.
[8] Y. S., Cho, S. C., Moon, S., Jeong, I., Oh, , K., Ho Ryu , "Clustering Method Using Item Preference Based on RFM for Recommendation System in U-Commerce", Ubiquitous Information Technologies and Applications, Vol. 214, 2013, pp.353-362.
[9] Z., khademali, A., harounabadi, J., mirabedini, "A new intelligent algorithm to creat a profile for user based on web interaction", Journal of management science letters, vol. 3, no.4, 2013, pp. 1155-1160.