Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

A new intelligent algorithm to create a profile for user based on web interactions

Zeinab khademali[a*], Ali Harounabadi[b] and Javad mirabedini[b]

[a]*Department of Computer Science, Islamic Azad University, Dezfoul Branch, Dezfoul, Iran*
[b]*Department of Computer Science, Islamic Azad University, Central Tehran Branch, Tehran, Iran*

CHRONICLE

Article history:
Received December 2, 2012
Received in revised format
2 March 2013
Accepted 5 March 2013
Available online
March 9 2013

Keywords:
Web mining
Clustering
Recommender engine
Web usage mining
Neural network

ABSTRACT

This paper presents a method to classify the web user's navigation patterns automatically. The proposed model of this paper classifies user's navigation patterns and predicts his/her upcoming requirements. To create users' profile, a new method is introduced by recording user's settings active and user's similarity measurement with neighboring users. The proposed model is capable of creating the profile implicitly. Besides, it updates the profile based on created changes. In fact, we try to improve the function of recommender engine using user's navigation patterns and clustering. The method is based on user's navigation patterns and is able to present the result of recommender engine based on user's requirement and interest. In addition, this method has the ability to help customize websites, more efficiently.

1. Introduction

Web-usage mining has become an area of extensive investigation (Kosala & Blockeel, 2000; Berendt et al., 2001). However, the capability of Web-usage mining outcomes depends on the appropriate preparation of the input datasets. More specifically, errors in building the sessions and incomplete tracking users' personal activities in a site can yield in invalid patterns and baseless conclusions. There are different studies in this regard and Spiliopoulou et al. (2003), for instance, made an assessment on the performance of heuristics employed to rebuild sessions from the server log data. They presented a set of performance figures, which were sensitive to two kinds of reconstruction errors and appropriate for various knowledge discovery (KDD) applications.

Corresponding author.
E-mail: zeinab.khademali@gmail.com (Z. Khademali)

doi: 10.5267/j.msl.2013.03.010

According to Phatak and Mulvaney (2002), web access from mobile devices expresses its own unique challenges because of the existing resource constraints on the mobile devices such as power, form factor, bandwidth, Phatak and Mulvaney (2002) recommend to try and predict a user's actions instead of reacting to a user's requests.

Adomavicius and Tuzhilin (2005) performed a comprehensive review on the field of recommender systems and explained the current generation of recommendation techniques, which are normally categorized into three main categories including content-based, collaborative, and hybrid recommendation techniques. They also explained different limitations of current recommendation techniques and explained possible extensions, which could contribute to recommendation capabilities and make recommender systems more useful to an even wider range of implications. These extensions incorporate an improvement of recognizing users' requirements and include contextual data into the recommendation process. In addition, they could provide support for multicriteria ratings, and a provision of more flexible and less intrusive kinds of recommendations.

Nicholas et al. (2006), in other survey, investigated different related works on deep log analysis (DLA) reporting on the information seeking methods of users. Nicholas et al. (2005) demonstrated a powerful and new DLA technique for mapping and evaluating information seeking behavior. Nicholas et al. (2006) used DLA techniques to show what usage data can reveal information seeking behavior of virtual scholars – academics, and researchers. Mobasher et al. (1999; 2001) presented some scalable techniques for Web personalization based on association rule discovery from usage data. We explained that the method could reach better recommendation effectiveness through detailed experimental evaluation on real usage data.

Lin et al. (2000) investigated the implementation of association rule mining as an underlying method for collaborative recommender systems and reported that such method was inefficient for collaborative recommendation since they include different rules, which are not relevant to users. Cooley et al. (1999) presented data propagation techniques to identify unique users and user session. Breeding (2005) described different methods to streamline and optimize how a Web site works to improve both its visibility and usability. They study also explained how to analyze logs and other system data to compute the effectiveness of the Web site design and search engine. Finally, Forsati and Meybodi (2009) presented an algorithmic based on structure pages and user's usage information for recommending web pages.

2. The preliminary requirements and definitions

In this section, we the necessary assumptions for web mining, customizing based on web usage mining. We also present the method of clustering and neural network and the concepts.

2.1. Web mining

Web mining uses the idea of data mining technique to find necessary data among documents and web services. Web structure mining in another idea, which analyzes nodes and structural relationships in a website based on models represented in graphs.Web content mining is also another process, which deals with discovering necessary information from texts, images, voice and visual data through web. Another term is associated with web usage mining, which is a process concentrates on techniques capable of predicting user's behavior interacting in web. The main functions in web usage mining are to retrieve comprehensive data from profile storage and using web servers based on user's browse.This process itself is categorized into three parts including pre-processing, pattern discovery and pattern analysis.

2.2. Clustering

The collection of input models $X= \{x_1, x_2,\dots, x_n\}$ includes 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 $C=\{C_1 , C_2 , \dots , C_k \}$, which do not overlap with each other. In this paper, *k*-means algorithm is used to cluster similar users, which is a popular technique for many clustering applications.

2.3. Neural network

Artificial neural network is an idea for processing some data, which are inspired from biological neural network and it acts like a brain. This system includes large numbers of processing elements called neuron, which act harmoniously to solve the problem. The distinctive advantage of these networks is their excessive capability along with simplicity.

3. The proposed method

Data recovery often arrives along with error since the available profiles in a server, saved sub sequentially, do not belong only to a user but they are available for various components. In addition, there are various search information kept for each user as well and these data must be pre-processed and prepared before implementation. Processing web logs incorporates data cleaning, user as well as user session identification. After preparing data and identifying users and their sessions, we build session vector as follows,

User's session can be described in terms of a vector of weight of page views during a specific period since a session includes all activities performed by the users from their arrival to site until their departure. A threshold is taken into account for the session duration and it this duration excesses from certain predefined level, it is a sign of the other access session of user. Based on this experiment , a thirty minute threshold is suggested for session duration (Berendt et al., 2011; Spiliopoulou et al., 2003). User session is also expressed in the following way: Let *p* be a collection of all accessible pages by site users with $p= \{p_{1,}\ p_2,\ \ldots,\ p_M\}$ provided that each p_i be distinguished by a particular URL. The collection of *S* also shows a subset of access sessions of users provided that each S_i be a subset of *P*.

$S= \{s_1, s_2,\ldots,s_n\}$

Each session is an *M*-dimensional vector as follows,

$S_i=\{W(P_1,S_i),W(P_2,S_i),\ldots,W(P_m,S_i)\}$

where the weight of each page, p_j, is determined in *i* session and every page weight shows the amount of user's interest to that page. In fact, to determine weight and amount of user's interest to the page, two factors of frequency and duration of page must be considered as follows,

$$\text{frequency(page)} = \frac{Number\ of\ visits(page)}{\sum_{page\in visited\ pages}(Number\ of\ visits(page))}, \tag{1}$$

$$\text{Duration(page)} = \frac{Total\ Duration(page)/Lenght(page)}{Max_{page\in visited\ pages}\ (Total\ Duration(page)/Lenght(page))}. \tag{2}$$

The relative importance of whole page is calculated by compounding two mentioned criteria. In this system, we use from the harmonious average of frequency and duration to explain the amount of user's interest to a web page in a session like below:

$$\text{Interest(page)} = \frac{2 * Ferequency(page) * Duration(page)}{Ferequency(page) + Duration(page)}. \quad (3)$$

Now we have a vector for every session where W_i determines the weight of the page i in a particular session. As the number of M dimension should not exceed from pre-specified number, the pages whose the amount of their support is high or low should be discarded.

3.1. Creating user's profile

Each user has k sessions such that $S_1, S_2, \ldots, S_k$ are collection of i user sessions. Average vector of S_{ui} is considered as a criterion or *ui* user interest. Weight of each page in average vector obtains from the average weight of that page in all user sessions. To achieve more efficient results, in addition to history of user's behavior, his/her trivial session can be used as well.

3.2. Clustering profiles

Now, the vector of average sessions needs to be compared with each other and they need to be clustered based on their similarities. In this algorithm, the number of clusters must be entered into an algorithm as an input parameter and cosines distance should be implemented to calculate the distance between two objects. The collection of clusters is as follows,

$C=\{c_1,c_2,\ldots,c_K\}$

As a representative for each cluster, we obtain the average of each m_c cluster, which shows the user navigation pattern of each cluster in a particular collection of accessible web pages. At last, as the result of profile clustering, there will be a collection like the following,

$NP=\{np_1,np_2,\ldots,np_K\}$

where each p_i is a subset of web page collection p. After training neural network, by entrance of new user to site, we need to prepare current user session in such a way that it would be possible to enter to neural network. Now it should be determined that the profile of current session belongs to which navigation pattern. Then, the profile of current session is given to the entrance of neural network and the network will determine appropriate cluster for the session. When the number of cluster is determined, pages of unvisited cluster in current session have high potentials to be next visiting candidate page and they will be kept into suggested list.

4. Results and Conclusion

In this paper, to provide useful and required data for users, a new method was introduced on the basis of user navigation patterns, which is capable of obtaining results by recommender's engine based on user's requirement. The preliminary results indicate that the proposed model of this paper performs better than alternative methods. The proposed method separates the pages, which are relevant to user's interest from irrelevant ones and to examine the impact of new method, the researcher performed a survey on the structure of user's profile based on the history of their behavior. If in adjustable research for each user, the researcher concentrates on user's current session more than his/her. Search history will lead to efficient results. This system uses neural network to determine classification of user similarity and common interests. Fig. 1 and Fig. 2 summarize the results obtained for precision and coverage:

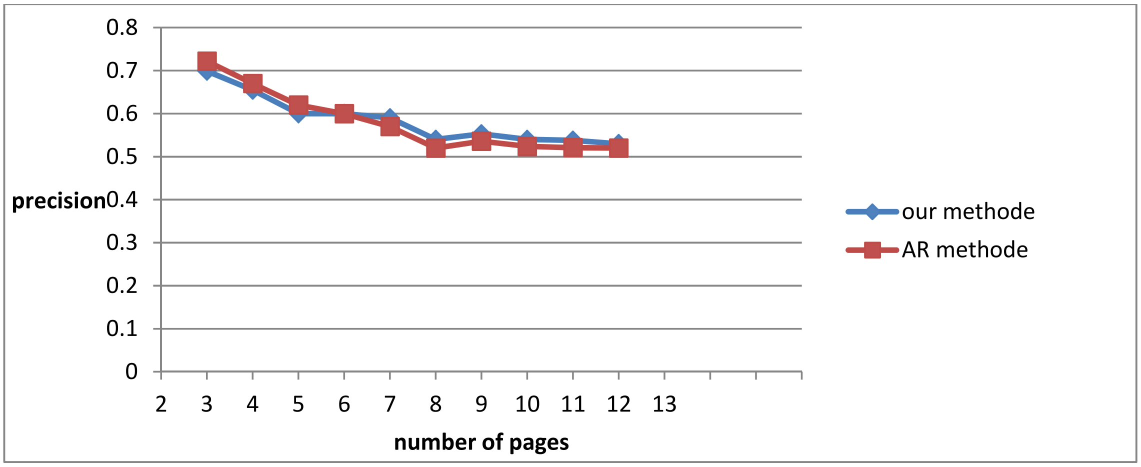

Fig.1. Precision of the recommendations

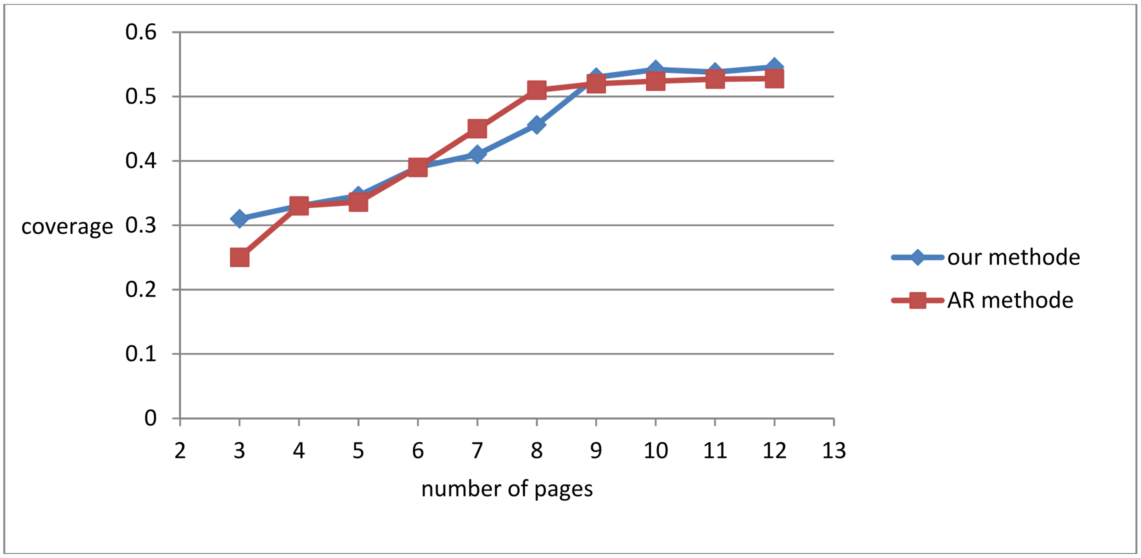

Fig. 2. Coverage of the recommendations

As explained earlier, the suggested method of this survey emphasizes on the structure of user's profile. As upcoming activities, we try to consider similar criteria in obtained clusters to be able to calculate the quality of suggestions provided by other users. In addition, we want to attribute the users to several clusters (overlapped clusters) and use the clusters for suggestion since people usually have different interest in real occasions.

Acknowledgment

The authors would like to give their special thanks to Dr Ali Hruon Abadi and Dr. Javad Mirabedini for their contributions on this work.

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(6), 734-749.

Breeding, M. (2005). Analyzing web server logs to improve a site's usage. The Systems Librarian. *Computers in Libraries*, *25*(9), 26-28.

Berendt, B., Mobasher, B., Spiliopoulou, M., & Wiltshire, J. (2001). Measuring the accuracy of sessionizers for web usage analysis. In *Workshop on Web Mining at the First SIAM International Conference on Data Mining* (pp. 7-14).

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*, *1*(1), 5-32.

Forsati, R., Meybodi, M.R. (2009). Algorithmic based on structure pages and user's usage information for recommending web pages. *The 2th Iran data mining conference*.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, *2*(1), 1-15.

Lin, W., Alvarez, S. A., & Ruiz, C. (2000). Collaborative recommendation via adaptive association rule mining. In *Proceedings of the International Workshop on Web Mining for E-Commerce (WEBKDD*.

Mobasher, B. (1999). A Web personalization engine based on user transaction clustering. In *Proceedings of the 9th Workshop on Information Technologies and Systems* (Vol. 18).

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* (pp. 9-15). ACM.

Mohamadi Dostdar, H., Forsati, R., & Meybodi, M.R. (2011) Recommender system of combined web based on 2-layers graph and partition of graph, *The 5th Iran data mining conference*.

Nicholas, D., Huntington, P., & Watkinson, A. (2005). Scholarly journal usage: the results of deep log analysis. *Journal of documentation*, *61*(2), 248-280.

Nicholas, D., Huntington, P., Jamali, H. R., & Tenopir, C. (2006). Finding information in (very large) digital libraries: a deep log approach to determining differences in use according to method of access. *The Journal of academic librarianship*, *32*(2), 119-126.

Nicholas, D., Huntington, P., Jamali, H. R., & Watkinson, A. (2006). The information seeking behaviour of the users of digital scholarly journals.*Information Processing & Management*, *42*(5), 1345-1365.

Phatak, D. S., & Mulvaney, R. (2002). Clustering for personalized mobile web usage. In *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on* (Vol. 1, pp. 705-710). IEEE.

Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing*, *15*(2), 171-190.