A New Method to Create the Profile and Improving the Queries in Web

Hale Falakshahi
Islamic Azad University
Science and Research Branch
Dept. of Computer Engineering
Neyshabur, Iran

Ali Harounabadi
Islamic Azad University
Central Tehran Branch
Dept. of Computer Engineering
Tehran, Iran

Majid Mazinani
Islamic Azad University
Science and Research Branch
Dept. of Computer Engineering
Neyshabur, Iran

ABSTRACT
Finding needed information among the existing information on the web can be very time consuming and difficult. To tackle this problem, web personalization systems have been proposed that adapt the contents and services of web sites based on the users' interests. Studying users' behaviors in the past with web usage mining techniques utilization can be worthy help in personalization affair. Web servers log files considered as a rich resource for finding users' behavioral patterns.

In this paper, users' behavioral patterns are obtained from studying of users' access and web usage mining utilization, especially users clustering. One of the innovative aspects of the research is selecting some behavioral features from users. These features include the 'pages view', 'page view frequency', 'time period of viewing the pages' and ' order of viewing the pages' which are stored in users' profiles. In addition is considered weight criterion for the first three features. Thus clustering has been done by considering this criterion with K-means algorithm. Neural network usage is another feature of proposed system to form recommender engine, which its function is to find proper behavioral pattern for users' session and forecast upcoming demands. As research conclusion presents recommender engine has the appropriate accuracy in prediction of user's inquiry.

Keywords
Web personalization, Web usage mining, Clustering, Neural network

1. INTRODUCTION
All aspects of the Web are developing rapidly. The Web is mainly used in sharing information in the early days, but now Web applications have spread to e-commerce, online games and other fields. At the same time, it needs higher requirement for website designs and functionality: improving the site structure for users to let them find the needed information quickly and accurately, recommending users the page or product that they are interested to provide personalized service, finding potential visitor groups, making accurate market positioning for different users [1].Therefore, a traditional technology of data mining has been used in Web mining. The Web server log has a perfect structure that it contains users' browsing behaviors which can reveal a wealth of information and provide a good prerequisite for the Web mining [2]. Web mining is categorized into three active research areas according to what part of web data is mined, of which web usage mining, also known as web-log mining, which its function is to study user access information from logged server data in order to extract interesting usage patterns [3].

There has been an increased demand for understanding of web-users due to the Web development. Based on different criteria, users can be clustered and useful knowledge can be extracted from web user access patterns. The purpose of finding similar interests among the web-users is to discover knowledge from the user profile. The definition of the similarity is application dependent. The similarity function could be based on visiting the same or similar pages, or the frequency of access to a page [4], or even on the visiting orders of links. In the latter case, two users that access the same pages could be mapped into different groups of interest similarities if they access pages in distinct visiting orders. In this paper, several similarity measures are proposed to capture the users' interests. K-means algorithm is then developed to cluster web users such that the users in the same cluster are closely related with respect to the similarity measure. Neural network usage is another feature of this system to form recommender engine, which its function is to find proper behavioral pattern for users' session and forecast upcoming demands. The rest of the paper is organized as follows: in section 2, related work will be introduced. Proposed method will be explained in section 3. The experimental results for evaluating proposed approach are presented in section 4.

Finally, Section 5 concludes the paper.

2. RELATED WORK
Recently, Web Usage Mining (WUM) is an active area of research which is beneficial for web personalization [3]. Existing web usage mining techniques include statistical analysis [5], association rules [6], sequential patterns [7], classification [8], and clustering [9]. An important topic in WUM is clustering web users. By analyzing the characteristics of the clusters, web users can be understood better and thus can be provided with more customized services. Clustering allows us to group clients or data items that have similar characteristics. Some approaches to clustering analysis have been developed for mining the Web server logs. Perkowitz and Etzioni [10] discuss adaptive Web sites that learn from user access patterns. The PageGather algorithm uses the page co-occurrence frequencies to find clusters of related but unlinked pages. Mobasher, Cooley and Srivastava [9] propose a technique for capturing common user profiles based on association-rule discovery and usage-based clustering. Cooley [11] introduces an algorithm that classifies users using a hypergraph partitioning technique. Cooley's method is used to identify particularly interesting and similar path histories, but it cannot be used to gain an overall picture

of all usage of a Web site. Nasraoui and Krishnapuram [12] use unsupervised robust multi-resolution clustering techniques to discover Web user groups. Xie and Phoha [13] use belief functions to cluster Web site users. They separate users into different groups and find a common access pattern for each group of users. Xu and Liu [3] cluster web users with Kmeans algorithm based on web user log data; they introduced 'hits' concept, hits mean one kind of user browsing information. The hits of all users who access the Web pages of a Web site can be extracted during a given period of time, $\text{hits}(i, j)$ is the count of user i accesses Web page j during a defined period of time. Count of visiting the pages is the criterion that is used for clustering. Petridou et al [14] emphasize the need to discover similarities in users' accessing behavior with respect to the time locality of their navigational acts. In this context, they present two time-aware clustering approaches for tuning and binding the page and time visiting criteria. The two tracks of the proposed algorithms define clusters with users that show similar visiting behavior at the same time period, by varying the priority given to page or time visiting.

3. PROPOSED METHOD

The method presented in this paper is based on web mining of server logs. It begins with preprocessing of server logs and then users' sessions are extracted. Four features of user behavior will be presented that amount of them are stored in user profile. Then the methods of comparing similarity between users based on these features are expressed. Weight criterion for the first three features is considered because the importance of these behavioral characteristics is dependent on the type and content of each site. User clustering will be done by K-means algorithm. The output of this stage is the clusters of users that users in each cluster have most similar behavior together. Neural network usage is another feature of this system to form recommender engine, which its function is to find proper behavioral pattern for users' session and forecast upcoming demands. The steps of proposed method are depicted in figure 1. In the following each of them will be described in details.

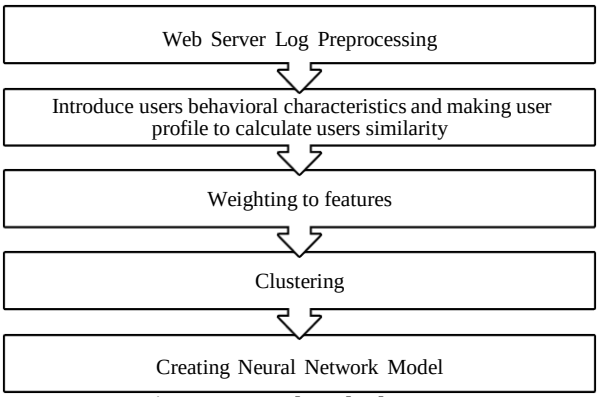

Fig 1: Proposed method steps

3.1 Web Server Log Preprocessing

Preprocessing of web server log files is conducted to identify user sessions. Web servers often registered all the activities of users in the form of web server logs. Because of the different configurations of servers, there are several types of server logs. But normally server log files, have the same basic information, such as client IP address, time of request, requested URL, the status code of HTTP, references and more. Several preprocessing operations should perform before applying the web usage mining techniques on the web server logs. These operations in the scope of this research include data cleansing and identifying and separating users' sessions. All data in web server logs are not suitable for web usage mining. So to remove the improper data from the log file, data cleansing step is accomplished by including the following data is deleted:

- Requests that are performed by automated programs such as spiders and web crawler data should be removed.
- Requests that are not completed. For example, the requests are faced with the HTTP error.
- Requests have response other than Get and Post.

A user session is a set of pages seen by the user during a special visit from a website. Before applying web usage mining techniques, web server logs should be grouped into meaningful sessions. In this study pages are considered as a session that are requested in a period of time less than equal to a certain time (e.g. 30 minutes).

3.2 Introduce Users Behavioral Characteristics and making user profile to Calculate Users Similarity

Collecting the appropriate features of user which represent user's behaviors and placed in web usage mining scope is the main goal of this paper. These features will be placed in the users' profiles. In the following these features will be introduced:

3.2.1 Pages View

One way that can be helpful in understanding user's interests is to consider what pages of site is almost viewed by user. User typically visits the pages that satisfy his or her information needs, so page view factor can help to understand user's interests and information needs. Therefore, is assumed that in the website which the aim is to personalize it, there are m users that viewed some of n pages of website. Set U is defined as all users who visit the website $U = \{u_1, u_2, \dots, u_m\}$ and set P is defined as all pages of web site $P = \{p_1, p_2, \dots, p_n\}$. Then $visit(u_i, p_j)$ is defined, if page p_j is accessed by user u_i the amount of this will be 1 and otherwise is 0. This vector can be obtained by retrieving the access logs of the site and will be stored in user profile. If two users accessed the same pages, they might have some similar interests in the sense that they are interested in the same information (e.g., news, products etc). The number of common pages they accessed can measure this similarity. The cosine similarity is used as the similarity measure. The measure is defined by

$$Sim_v(u_i, u_j) = \frac{\sum_k (visit\ (u_i, p_k) * visit\ (u_j, p_k))}{\sqrt{\sum_k visit\ (u_i, p_k)^2 * \sum_k visit\ (u_j, p_k)^2}} \quad (1)$$

Where $Sim_v\ (u_i, u_j)$ is a similarity between user u_i and user u_j based on viewing the same pages.

3.2.2 Page View Frequency

Another feature that can help in identifying user preferences is the number of times a user returns to a page during a period of time and visit the page again. Whatever the number of times of accessing to one page is frequent the user is more likely to be interested in that page. This feature is stored in user profile

in form of vector $count\ (u_i, p_j)$ shows the number of time the user u_i accessed to the page p_j. The similarity between two users can be measured by counting the number of times they access the common pages. The cosine similarity is used as the similarity measure. In this case, the measure is defined by

$$Sim_c\ (u_i, u_j) = \frac{\sum_k (count\ (u_i, p_k) * count\ (u_j, p_k))}{\sqrt{\sum_k count\ (u_i, p_k)^2 * \sum_k count\ (u_j, p_k)^2}} \quad (2)$$

Where $Sim_c\ (u_i, u_j)$ is a similarity between user u_i and user u_j based on page view frequency.

3.2.3 Time Period of Viewing the Pages

The time user spends on viewing the page can show her or his interest on it. This feature is stored in user profile in form of vector $time\ (u_i, p_j)$ that shows the amount of time user u_i spend on viewing the page p_j.The similarity between two users can be measured more precisely by taking into account the actual time the users spent on viewing each web page. The cosine similarity is used as the similarity measure. In this case, the measure is defined by

$$Sim_t\ (u_i, u_j) = \frac{\sum_k (time\ (u_i, p_k) * time\ (u_j, p_k))}{\sqrt{\sum_k time\ (u_i, p_k)^2 * \sum_k time\ (u_j, p_k)^2}} \quad (3)$$

Where $Sim_t\ (u_i, u_j)$ is a similarity between user u_i and user u_j based on time period of viewing the page.

3.2.4 Order of Viewing the Pages

In this case, two users are considered having the same interests only when they access a sequence of web pages in the same order. The similarity between users, in such a situation, can be measured by checking the access orders of web pages in their navigation paths. Order of viewing the pages is needed for each user. Therefore, is assumed that Q = $q_1, q_2, \dots\ q_r$ is a navigation path for one user in length of r where q_i ($1 \le i \le r$) shows a user access to a page in i-th order. Q_l is defined as set of Q sub-paths:

$$Q_l = \{q_i, q_{i+1}, \dots, q_{i+l+1} \mid i = 1, 2, \dots, r - l + 1\} \quad (4)$$

Q_l contains all the sub-path with length of l in Q. It is obvious l is less than or equal to r. For each user all elements of Q_l ($1 \le l \le r$) is computed. Let Q^i be a path order of visiting of user u_i. To calculate the similarity between users the number of viewing paths by the user is not enough because the paths with less length are likely to be repeated. That is why after calculating the number of viewing paths by a user is multiplied to the lengths of these paths. Finally the formula for order of viewing the pages feature is defined by:

$$\mathrm{Order}(u_i, \vec{p}) = v(u_i, \vec{p}) \times l(\vec{p}) \quad (5)$$

$\vec{p}$ is a sub-path. For all the available paths $order(u_i, \vec{p})$ are computed and stored in user profile. $\mathrm{Order}(u_i, \vec{p})$ is weight frequency which user u_i has passed path $\vec{p}$. $v(u_i, \vec{p})$ shows the times that path $\vec{p}$ is viewed by a user. The length of this path is considered as $l(\vec{p})$.

The cosine similarity is used as the similarity measure between two users. In this case, the measure is defined by

$$Sim_o(u_i, u_j) = \frac{\sum_{(\vec{p} \in Q^i \cup Q^j)} (order(u_i, \vec{p}) * order(u_j, \vec{p}))}{\sqrt{\sum_{(\vec{p} \in Q^i)} order(u_i, \vec{p})^2 * \sum_{(\vec{p} \in Q^j)} order(u_j, \vec{p})^2}} \quad (6)$$

Where $Sim_o\ (u_i, u_j)$ is a similarity between user u_i and user u_j based on order of viewing the pages.

3.3 Weighting Features

As the importance of mentioned features depends on the site genre, weight criterion is considered for the first three features. For first three features user- page matrix is gained but for the last feature user-sub path matrix is gained so in weighting criterion 'Order of Viewing the Pages' feature is not included and it will be considered separately. The weights are experimental values and the optimal values for them will be achieved during the site analysis and evaluation over time. $w(u_i, p_j)$ is considered as weighting function for first three features and defined as follows

$$w(u_i, p_j) = a\ visit(u_i, p_j) + b\ count(u_i, p_j) + c\ time(u_i, p_j) \quad (7)$$

Vector $w(u_i, p_j)$ is a weight given to page p_j based on user u_i features and a, b and c are experience values in the site. The similarity between users can also be gained based on weighting criterion by use of cosine similarity.

3.4 Clustering

In this stage K-means clustering algorithm is performed. The flow of algorithm is shown as the following steps:

(1) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

(2) Assign each object to the group that has the closest centroid.

(3) When all objects have been assigned, recalculate the positions of the K centroids.

(4) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

At the end of clustering stage in proposed method *K* clusters will be gained that users in each cluster patterns would be most similar to each other with respect to individual preferences.

3.5 Creating Neural Network Model

This component of system duty is to get current user session and generate a list of suitable offer. To find the most similar cluster to current user session, neural network is used, then appropriate pages is offered to the current user based on the users' behavior of this cluster. After training the neural network, the current user profile given as input to network the output would be the number of the cluster that are most similar to current user behaviors. When suitable cluster was determined for current user, those pages would be placed in offer list which are existed in this cluster but have not seen by current user.

4. EXPERIMENTAL EVALUATION

The real data set were used for this study which related to Palood[1] dairy products company. The result of this report is

[1] www.palooddairy.com

for one month (July 2013).The size of access log for a month is 25 MB.

The information contained in the user access log of this site includes more details of users' requests so unnecessary information has been refined. The information stored for this study include: requesting IP address, date and time of the request, Parameters sent in every web address, communication method, user's browser type and operating system and page sizes. After collecting information data preprocessing was done. The values of similarity functions were gained. Weight criterion was calculated and then given to RapidMiner software in form of user-page matrix for clustering using K-means algorithm. The values of a, b and c in equation (7) set to 0.3, 0.5 and 0.2 respectively. Then similarity based on "Order of Viewing the Pages" was gained and given to RapidMiner software in form of user-sub path matrix. K-means algorithm is applied to cluster web users with different k values, whose result is demonstrated in figure 2. As the figure 2 shows the optimal precision would be gained when number of clusters (K) is set to 6 or 7.

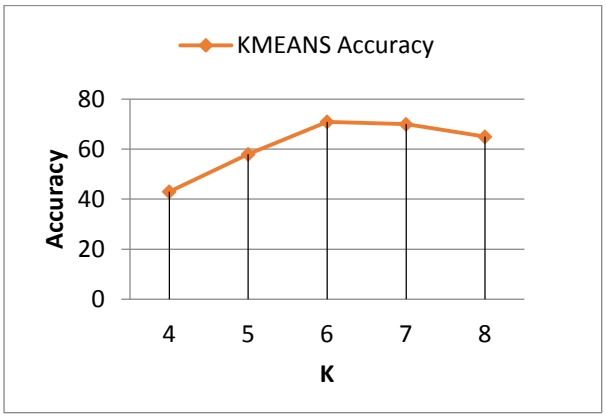

Fig 2: K-means accuracy with different K values

Neural network usage is another feature of this system to form recommender engine, which its function is to find proper behavioral pattern for users' session and forecast upcoming demands. Three-layer neural network was used. Neural network development needs partitioning the data into two categories: "training" and "test". The data for first two weeks was considered for training neural network and second two weeks used as test data. Sigmoid activation function was used in hidden layer and output layer. AutoMLP was used for learning algorithm. RapidMiner software parameters for neural network section were set as shown in table 1.

The most popular methods of evaluation are the precision and recall. Precision and recall values are between zero and one, and higher values indicate better way. In this paper precision and recall values were calculated for prediction based on weight criterion feature and order of viewing the pages feature. In figure 3 the precision of these two approaches is shown according to number of pages. Recall results is also shown in figure 4 according to number of pages. As shown in these two figures the prediction based on weight criterion has better results so it is compared with two methods were explained in related work section.

Table 1. Neural network setting parameters in RapidMiner software

Field name	value
Training cycles	2000
Learning rate	0.3
Momentum	0.2
Error epsilon	0.1 E -5

In the study [3] Xu and Liyu used the number of visiting the pages for clustering. Prediction based on their clustering method was done and their method in this paper is determined as 'hits'. Also paper method was compared with study [14] that explained in related work section. Petridou et al proposed two time-aware clustering approaches for tuning and binding the page and time visiting criteria. Paper method was compared with their best result in clustering section. Also their method is determined as 'time-aware'. Precision and Recall results are shown in figure 5 and 6 respectively. The paper proposed had a better result in comparing with them.

5. CONCLUSION

In this paper, in order to provide each user with more relevant information, a method has been presented for users clustering. Neural network usage is another feature of this system to form recommender engine. The proposed algorithm was applied to the real world data. Precision and recall were used for evaluation. The experimental results show the proposed algorithm has a better consequence compared to similar work done in the past.

6. REFERENCES

[1] Zheng, W., & Zhang, M. (2011, September). The investigation for Web user clustering based on interest. In Electronics, Communications and Control (ICECC), 2011 International Conference on (pp. 553-556). IEEE.

[2] Agosti, M., & Di Nunzio, G. M. (2007). Gathering and mining information from web log files. In *Digital Libraries: Research and Development* (pp. 104-113). Springer Berlin Heidelberg.

[3] Xu, J., & Liu, H. (2010, October). Web user clustering analysis based on KMeans algorithm. In *Information Networking and Automation (ICINA), 2010 International Conference on* (Vol. 2, pp. V2-6). IEEE.

[4] Yan, T. W., Jacobsen, M., Garcia-Molina, H., & Dayal, U. (1996). From user access patterns to dynamic hypertext linking. *Computer Networks and ISDN Systems*, *28*(7), 1007-1014.

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

[6] Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001, November). 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.

[7] Yang, Q., Zhang, H. H., & Li, T. (2001, August). Mining web logs for prediction models in WWW caching and prefetching. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 473-478). ACM.

[8] Li, I. T. Y., Yang, Q., & Wang, K. (2001, May). Classification Pruning for Web-request Prediction. In *WWW Posters*.

[9] Mobasher, B., Cooley, R., & Srivastava, J. (1999). Creating adaptive web sites through usage-based clustering of URLs. In *Knowledge and Data Engineering Exchange, 1999. (KDEX'99) Proceedings. 1999 Workshop on* (pp. 19-25). IEEE.

[10] Perkowitz, M., & Etzioni, O. (1998, July). Adaptive web sites: Automatically synthesizing web pages. In *AAAI/IAAI* (pp. 727-732).

[11] Cooley, R. W. (2000). Web usage mining: discovery and application of interesting patterns from web data (Doctoral dissertation, University of Minnesota).

[12] Nasraoui, O., & Krishnapuram, R. (2002). An evolutionary approach to mining robust multi-resolution web profiles and context sensitive URL Associations. *International Journal of Computational Intelligence and Applications*, *2*(03), 339-348.

[13] Xie, Y., & Phoha, V. V. (2001, October). Web user clustering from access log using belief function. In *Proceedings of the 1st international conference on Knowledge capture* (pp. 202-208). ACM.

[14] Petridou, S. G., Koutsonikola, V. A., Vakali, A. I., & Papadimitriou, G. I. (2008). Time-aware web users' clustering. *Knowledge and Data Engineering, IEEE Transactions on*, *20*(5), 653-667.

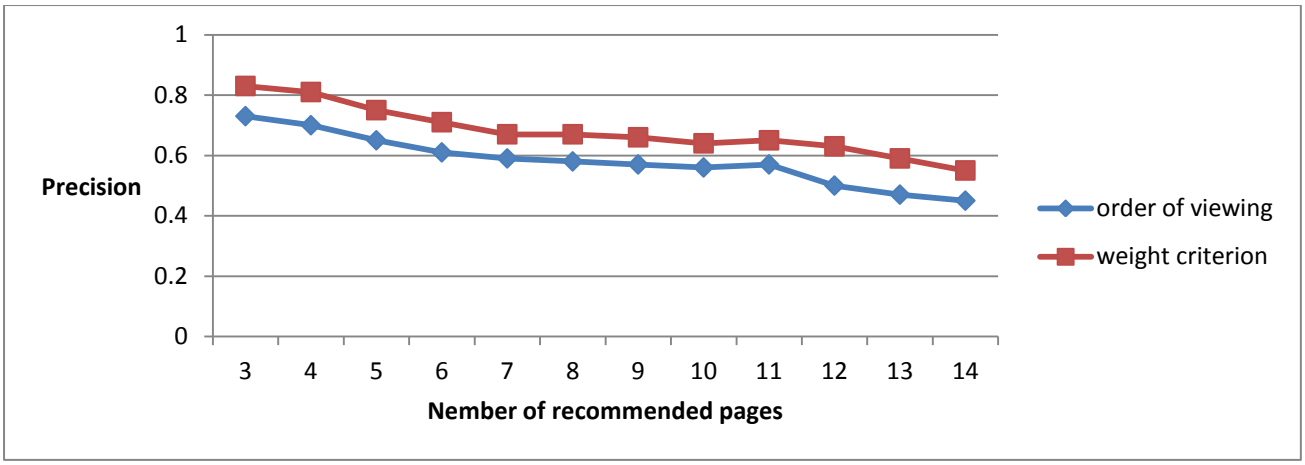

Fig 3: Precision obtained using clustering based on order of viewing the pages feature and weight criterion feature

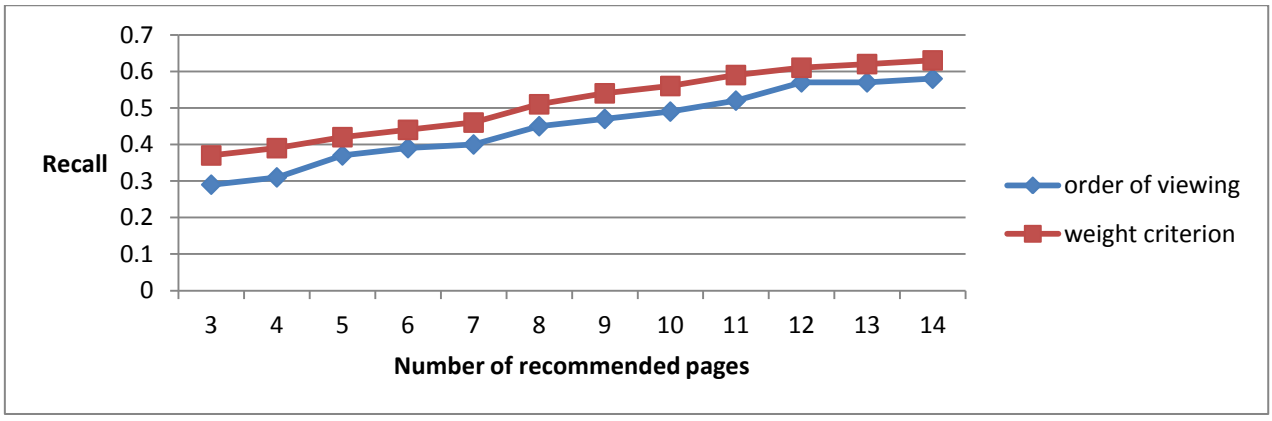

Fig 4: Recall obtained using clustering based on order of viewing the pages feature and weight criterion feature

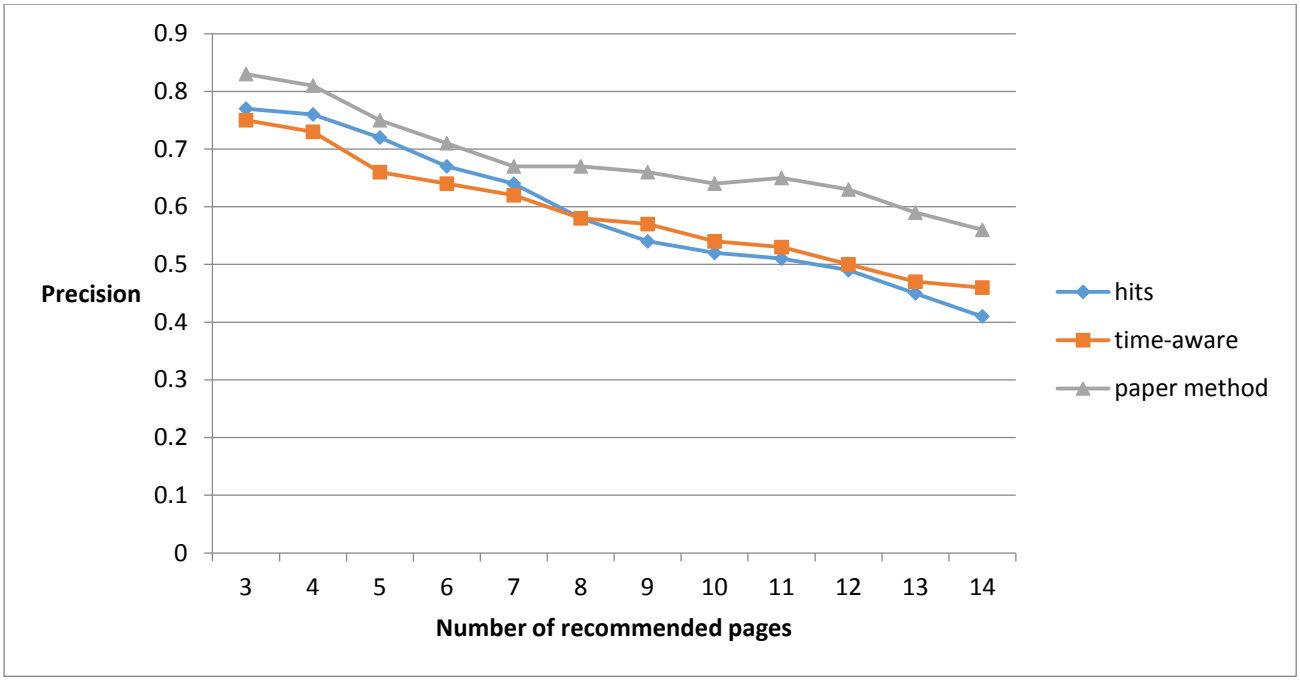

Fig 5: Comparison of the proposed method precision

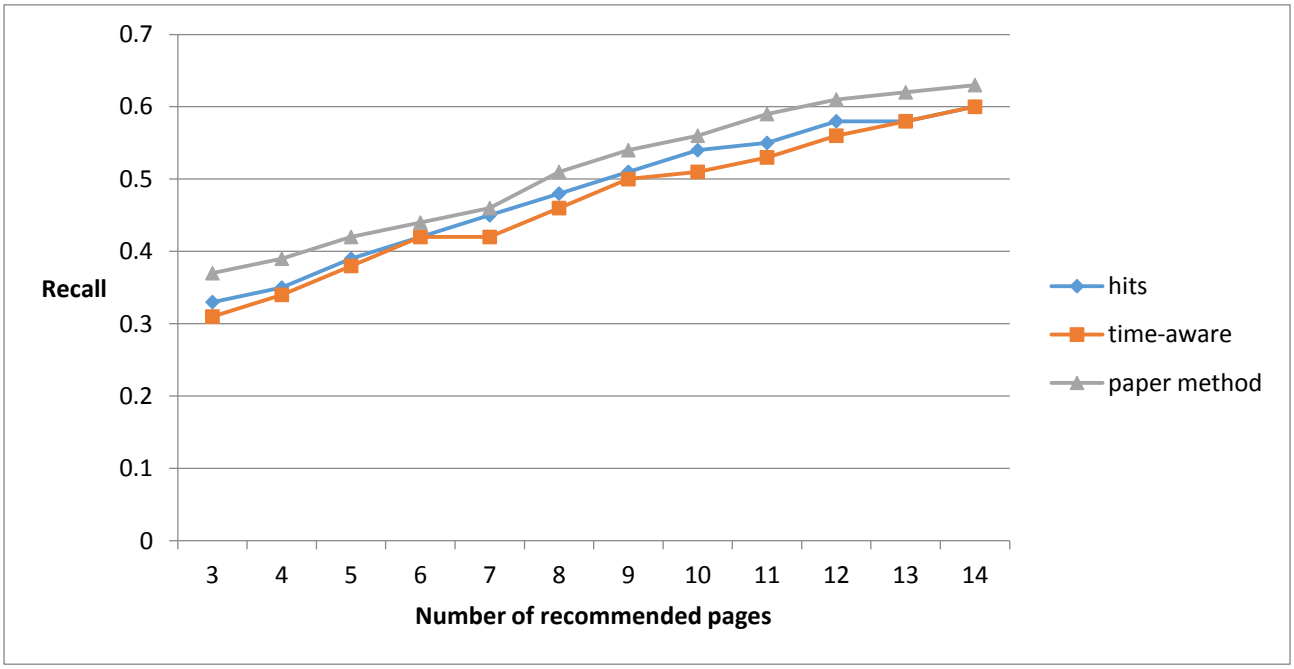

Fig 6: Comparison of the proposed method recall