I J C T A, 10(8), 2017, pp. 679-687 ISSN: 0974-5572
© International Science Press

Session based Collaborative Filtering for web page Recommender (SCFR) System based on clustering

Poornalatha G.* and Prakash S. Raghavendra**

ABSTRACT

With the explosive growth of the world wide web, getting relevant and useful information from vast quantity of data has become a significant problem. Web page recommender system assists user by providing recommendations to ease their navigation through a web site. Many recommender systems have been developed to discover web pages that may be useful to the user. This paper proposes a novel recommender system which adopts the concept of collaborative filtering. The performance of proposed recommender system is evaluated based on precision and recall metrics. Also, results obtained are encouraging in terms of precision and recall compared to couple of other results in the literature.

Index Terms: collaborative filtering, k-means, server log, session.

1. INTRODUCTION

The number of people who use internet is increasing day by day in recent era. The users are enforced to spend more time in looking for appropriate and desired information from huge quantity of web pages available over internet. Web page recommender system addresses this problem by providing suitable suggestions to user. The goal of recommender system is to determine the web pages relevant to user based on user's present action and the actions performed by other users earlier.

A recommender system plays an important role in various applications like e-commerce. Reference [1] analyzes various web sites and explained the use of recommender systems that increases the sales. Reference [2] integrates data mining techniques and social network techniques to analyze web server logs to assist web site owners to understand the visitor's behavior. This may help the site owner to decide placement of advertises for new products, announce discounts, give special offers, provide link to popular products etc. so that, the customers can take advantage of new schemes or offers which may not be known to customer otherwise. The owner will be benefitted from extra sales made through recommendations and, there is a chance of converting a person browsing the web into a customer. Thus, recommendation helps both customers as well as web site owners.

As people participate actively in social networking and peer-production sites, implicit relations may emerge from various activities [3]. Discovering such relations by mining the users' activities leads to better recommendations. Also, people depend on information available on the web for various activities like financial, health, career, and education regularly. Even search engines play a passive role by trying to provide relevant information that is explicitly asked or requested by the user. Many times, a user may not know existence or availability of some useful information, but the recommender system may suggest such useful information to the user implicitly based on past activities. Thus, recommender systems provide

* Information and Communication Technology Department, MIT, Manipal University, Manipal, *Email: Poornalatha.g@manipal.edu*

** National Institute of Technology Karnataka (NITK), Surathkal, Mangalore, *Email: srp1970@gmail.com*

relevant and previously unknown information to the user and hence play a significant role in many web-based applications.

Reference [4] represented user session as a vector, clustered sessions by using leader algorithm and dynamically suggested links to user. Reference [5] presents techniques to discover aggregate profiles that can be used by recommender systems for real time web personalization. The effective and scalable techniques for web personalization based on association rule discovery from usage data was proposed in [6]. Reference [7] proposes a web recommendation system based on the weighted association rule (WAR) model. They extended the association rule mining by assigning a significant weight to the pages based on time spent by each user on each page and visiting frequency of each page. Reference [8] mines indirect association rules for web recommendation. The relational fuzzy subtractive clustering was used as the first level modeling in [9] and then mined association rules within individual clusters. They proposed a two-level model-based technique, which is scalable and is an enhancement over association rule based recommender systems. Reference [10] proposes a recommender system based on fuzzy association rule mining. Reference [11] proposes algorithm for pruning sequence association. Using association rules for web page recommendation involves too many rules and difficult to find a suitable subset of rules to make accurate and reliable recommendations [12] and hence, there is a need to look at other alternative methods.

The other common methods used for recommender system to improve its performance are clustering [5], [13] “to be published [14]-[15], graph based [16], web content based [17]-[18] etc. Various hybrid models are also proposed to overcome the limitations of these individual methods. For example, [19] proposes a hybrid recommender system that combined results of several recommender techniques based on web usage mining. Reference [20] proposes a hybrid model that includes markov model, sequential association rule, association rule and a default model that recommends based on frequency from a whole data set. Reference [21] proposes a hybrid recommendation method based on the ant colony metaphor. The different algorithms (e.g. top N, sequence patterns, collaborative filtering) were combined in a single recommendation database in [22]. Though the intention of hybrid models is to overcome drawback of each of the individual model, they require more time for both off line as well as on line phase because of applying individual models sequentially.

Recommender systems are generally classified into collaborative filtering and content-based filtering [23]. Collaborative filtering is a well-known technology that uses past navigation behavior to generate web pages as recommendations to user [24]. Collaborative filtering uses the known preferences of a group of users to make recommendations of the unknown preferences for other users [25]. The conventional collaborative filtering method finds recommendations from the complete data base of user sessions that lead to scalability problem. Reference [26] discusses various collaborative filtering algorithms. The proposed method uses clustering as well as collaborative filtering technique to achieve better performance in terms of precision and recall. The proposed SCFR system finds recommendations within the nearest cluster and not the complete data base of user sessions. Thus, SCFR resolves scalability issue associated with the conventional collaborative filtering method. Also, the pages that are visited more than once in a session are removed since they do not contribute for recommendations. Thus, the objective of the present paper is to discuss the proposed recommender model based on session collaborative filtering that improves performance in terms of recall and precision.

The remaining part of this paper is organized as follows. Section II explains the proposed session based collaborative filtering recommender system. The results are discussed in section III. Section IV gives concluding remarks followed by the references at the end.

2. PROPOSED RECOMMENDER SYSTEM

The goal of the present work is to develop a simple recommender system based on user sessions clustering and collaborative filtering to improve the performance over couple of existing recommender models. This section explains the activities of off line and on line components used in the proposed recommender system.

2.1. Session based Collaborative Filtering Recommender System

Figure 1 depicts the session based collaborative filtering recommender system, that recommends web pages to an active user, based on the previous similar sessions. The system has offline and online components. The details of activities of these two components are given below.

Off line Component: The majority activities of off line component are like that of any other web usage mining system. The web server log maintains ip address, date, time, HTTP method, requested URL, response code, number of bytes transferred, referrer etc. for each request. The log file is parsed to extract essential fields like ip address, date, time and the requested URL. The data is cleaned by removing image files, response code with error and empty requested URL field. Unique URLs are identified from cleaned data and unique identities are assigned for each unique URL.

The user sessions are created based on ip address, date and time fields. A session represents the sequence of web pages viewed by a user within certain time duration (30 minutes). If a user views page 1, page 3, page 1, page 5 and page 6 in succession, the session is represented as P1, P3, P1, P5, P6. Once sessions are created, repeated pages are removed since they do not contribute for recommendations. For example, in the above session, page P1 occurs twice and after removing duplicate pages the session becomes P1, P3, P5, P6. Possible reasons for existence of repeated pages in a session are, click on back button of browser, referring the earlier visited page again etc.

Once sessions are created and duplicate pages are removed, they are clustered by using k-means algorithm as given in Algorithm 1.

Algorithm 1. K means for recommender system

```
Input: user sessions, number of clusters n
Output: set of clusters
Steps:
•   Select n user sessions randomly as initial centroids
•   Loop
      * Assign each session to the nearest cluster
      * Count frequency of each page for respective clusters
      * Select pages that are accessed more frequently (> 50%) and concatenate them to form new cluster
        centroids
      * Exit if there is no change in centroids or sessions remain in same cluster with new centroid
•   End loop
```

The algorithm 1 gives major steps of k-means, used to cluster web user sessions. Depending on the number of clusters, the initial cluster centroids are randomly selected from user sessions. Each user session is assigned to nearest cluster after measuring its distance with cluster centroids. For measuring the distance between any two sessions VLVD method [27] and cosine distance method are used and are discussed in the next section. Once the initial clusters are formed, the new cluster centroids are computed. To get new cluster centroids, the count of each page in the respective cluster is used. The page count is determined during assigning the session to the cluster. Therefore, at the end of clustering the count for each page is available. The frequency of each page is determined by dividing the page count by the total number of sessions in the respective cluster. The frequency value greater than 0.5 indicates, the page is accessed by more than 50% of sessions in the cluster.

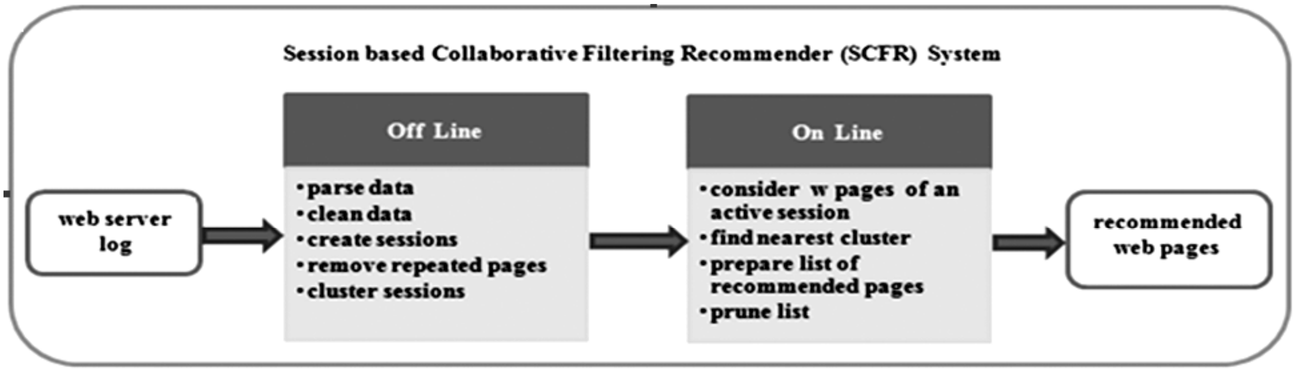

Figure 1: Session based collaborative filtering recommender system

Hence the page is considered as, one of the frequently viewed pages of the cluster. The pages that are accessed more frequently (>50%) are determined and are concatenated to obtain new cluster centroids. Thus, the new centroids represent the web pages that are accessed by more than 50% sessions in respective clusters.

On line Component: During the on line phase, first ‘w’ pages of an active user session are taken and is assigned to the nearest cluster. Once the nearest cluster is found, the distance between active session and other sessions in the cluster are computed. The sessions that are more like the active session are considered and list of pages that are not present in the active session is prepared. This page list is further pruned based on the frequency. Thus, top n pages are recommended from the cluster based on the frequency of co-occurrence in comparison with the active user session. The model is evaluated by using precision and recall as metrics.

3. EXPERIMENTAL ANALYSIS

This section describes the data sets used for evaluation of the proposed recommender system, discussed the results obtained and gives the comparison of the results obtained by the proposed recommender system with couple of other recommender systems.

Data Sets: The first set is NASA log taken from NASA Kennedy space center www server in Florida (http://ita.ee.lbl.gov/html/contrib/NASA-HTTP .html) which consists of more than 10,00,000 entries. The log has the data collected from 00:00:00 July 1, 1995 through 23:59:59 July 31, 1995, a total of 31 days. The data is parsed to extract the required attributes. The parsed data is cleaned to remove unwanted entries like image files like bmp, jpg etc., and script files like js. Entries with error code (other than 200) in response field are also removed. After processing the data, sessions are constructed based on ip address/URL field. Repeated pages from each session are removed to obtain sessions with unique pages.

The second set is Music Machine (MM) data set which consists of log for December 1997. The MM server log was obtained from http://www.cs.washington.edu/ai/adaptive- data/. The web site gives details of various musical instruments, samples, images, manufacturers of different types of instruments, details of dealers etc. The user sessions for the MM data set are created by preprocessing the log as described for the NASA data set.

Distance Methods: The distance methods used to find the similarities between any two user sessions for clustering is discussed in this section. The distance methods used are VLVD [27] and the cosine distance.

Cosine Similarity: Cosine similarity is the most common method used to find similarities between vectors in information retrieval. The cosine similarity between two sessions is given by equation (1), that uses dot product and magnitude to compute similarity between them. The resulting value ranges from 0 to

1 where 0 indicates the sessions are complete different and 1 results if sessions are exactly similar to each other. The steps used to find distance between two sessions is given in algorithm 2.

$$d_{\cos}\ \left(S_i, S_j\right) = \left(S_i.S_j\right) \Big/ \left(\|S_i\|.\|S_j\|\right) \tag{1}$$

Algorithm 2: cosine similarity between two user sessions

```
Input: two web user sessions $S_i$ and $S_j$
Output: distance d between $S_i$ and $S_j$
steps:
    • $l_1$ = number of pages in session $S_i$
    • $l_2$ = number of pages in session $S_j$
    • c = number of common pages accessed by sessions $S_i$ and $S_j$
    • $d_{COS}(S_i,S_j)$ = c / {(sqrt($l_1$) . sqrt($l_2$)}
```

Evaluation Metrics: The proposed SCFR system is evaluated by using precision and recall as metrics [28]. The precision is number of recommended pages that are relevant and the recall is number of pages that are correctly recommended. Let RP be the total number of pages recommended and n denote the length of an active session. First 'w' number of pages of an active session is used to assign it to the nearest cluster. The remaining part of an active session after first w pages is denoted by (n-w). Using the above given notations, precision and recall of a session are defined as given in equation (2) and (3) respectively.

$$precision\left(S_i\right) = \{RP \cap (n-w)\}/RP \tag{2}$$

$$recall\left(S_i\right) = \{RP \cap (n-w)\}/(n-w) \tag{3}$$

The 40% of sessions are considered as test set to evaluate the proposed recommender system. If TS is number of test sessions considered for evaluation, then overall or average precision and recall are given by equation (4) and (5) respectively.

$$precision\left(TS\right) = \sum_{i-1}^{TS} precision(S_i) \Big/ TS \tag{4}$$

$$recall\left(TS\right) = \sum_{i-1}^{TS} recall(S_i) \Big/ TS \tag{5}$$

Experimental Results: The MM data set is considered first for evaluating the proposed system with 5000 and 10000 sessions. In both the cases the total sessions are divided into 60:40 ratios. 60% of sessions are considered for building the recommender system as shown in the off line component of figure 1 and remaining 40% sessions are taken as test data. The average session length is 7.91 and 7.96 for 5000 and 10000 sessions respectively. For each test session, nearest cluster is found based on first 'w' pages and 'w' is assumed as 3. List of recommended pages is generated by comparing the 'w' pages of test session with other sessions of cluster based on cosine and VLVD similarity. The list with recommended pages is pruned based on frequency. The experiment is conducted for various size of recommended list ranging from top 1 to top 10. Figures 2, 3, 4 and 5 shows precision and recall values for 5000 and 10000 sessions, by using cosine and VLVD distance measures respectively. The recall increases as the number of recommended pages increases while the precision decreases as the number of recommended pages increases. This is because the average session length is 8 and first 3 pages are not considered for recommendation. The recommendation is given to the last 5 pages of session and maximum of 5 pages could be accessed by user though the recommendation list contains more than 5 pages. Therefore, the precision reduces as the recommendation list size increases. The results are almost consistent for both cases of cosine and VLVD distance measures as can be seen from figures 2, 3, 4 and 5. Though the results are almost same, it could be inferred that cosine similarity is not efficient than VLVD because of square root operation.

Similarly, NASA data set for the month of July is considered for experiment purpose. The proposed system is evaluated by using 5000, 10000 and 15000 sessions. The average session lengths of these sessions are 6.68, 6.77 and 6.82 respectively. Figure 6, 7, 8 and 9 shows recall and precision values for 5000, 10000 and 15000 sessions with VLVD and cosine similarity as a distance measure. Again, the results are consistent for various session sizes for NASA data set also. The recall is more than 60% for top 5 pages that indicates goodness of the proposed SCFR system.

A hybrid click-stream based collaborative model was proposed in [20]. They used four different existing models namely, markov model, sequential association rule, association rule and default model. The default model recommends pages based on frequency from whole data set. The experiment is done by applying these models in sequence with different combinations.

For example, SMAD hybrid model applies sequential association rule first to the active session. If sequential association rule does not cover the active session, then markov model is used. If markov model fails, then association rule is used. If association rule also fails, finally default model is used.

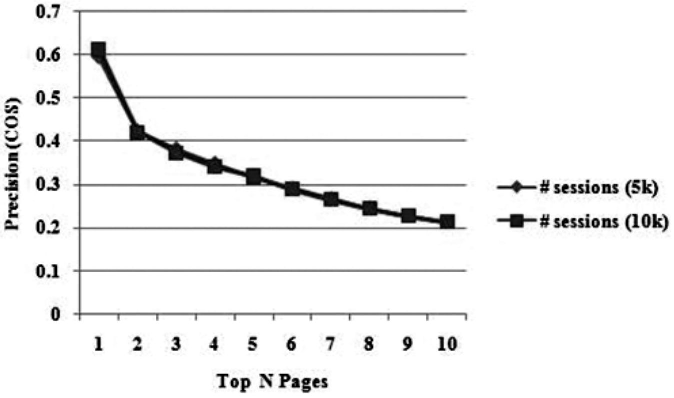

Figure 2: precision for 5k and 10k sessions with cosine similarity as distance measure

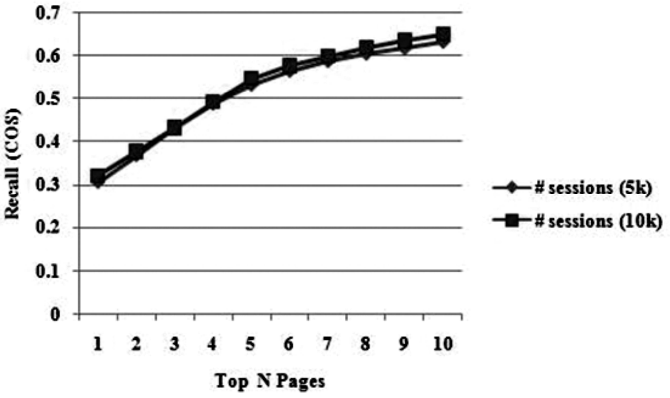

Figure 3: *recall* for 5k and 10k sessions with cosine similarity as distance measure

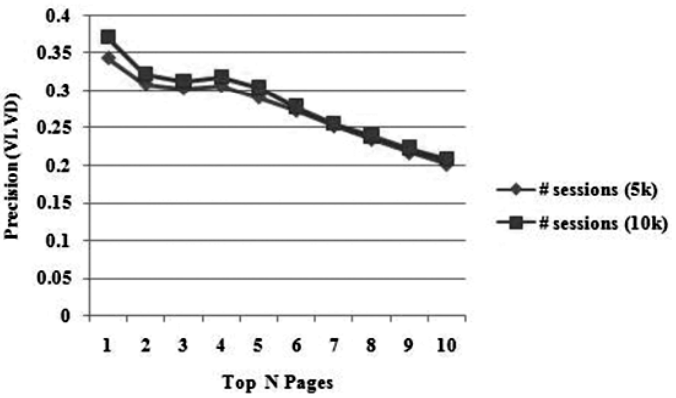

Figure 4: precision for 5k and 10k sessions with VLVD similarity as distance measure

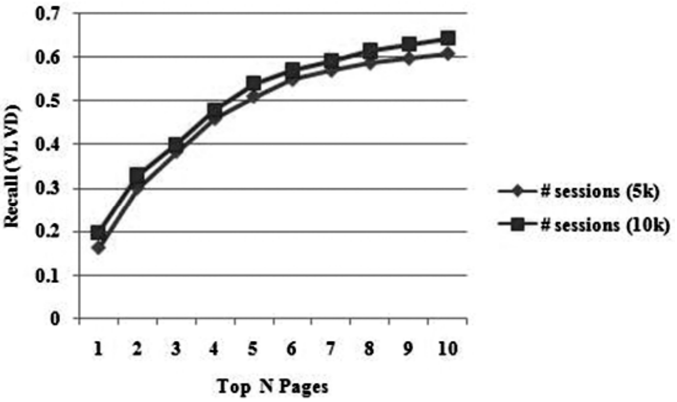

Figure 5: recall for 5k and 10k sessions with VLVD similarity as distance measure

Table 1
Recall of the various models for top n pages

	Top 2	*Top 4*	*Top 6*	*Top 8*	*Top 10*	*Top 12*	*Top 14*	*Top 16*	*Top 18*	*Top 20*
$\text{SCFRS}_{\text{VLVD}}$	0.3432	0.5131	0.5969	0.6424	0.6745	0.7018	0.7188	0.7338	0.7493	&.7619
$\text{SCFRS}_{\text{COS}}$	0.3454	0.5064	0.5894	0.6368	0.6673	0.6960	0.7143	0.7287	0.7440	0.7557
SMAD	0.0542	0.0892	0.1227	0.1483	0.1710	0.1920	0.2089	0.2244	0.2373	0.2516
ASMD	0.0445	0.0764	0.1059	0.1266	0.1453	0.1675	0.1861	0.2056	0.2234	0.2370
MASD	0.0536	0.0867	0.1199	0.1423	0.1623	0.1847	0.2037	0.2211	0.2381	0.2516

Table 2
Precision of the various models for top n pages

	Top 2	*Top 4*	*Top 6*	*Top 8*	*Top 10*	*Top 12*	*Top 14*	*Top 16*	*Top 18*	*Top 20*
$\text{SCFRS}_{\text{VLVD}}$	.2777	.2979	.2647	.2296	.2039	.1853	.1673	.1534	.1428	.1338
$\text{SCFRS}_{\text{COS}}$	0.2978	0.3003	0.2656	0.2315	0.2041	.1858	0.1690	0.1552	0.1444	.1355
SMAD	0.2186	0.1797	0.1648	0.1494	0.1378	.1290	0.1203	0.1131	0.1063	.1014
ASMD	0.1794	0.1540	0.1423	0.1276	0.1171	.1125	0.1072	0.1036	0.1000	.0955
MASD	0.2159	0.1748	0.1610	0.1434	0.1309	.1240	0.1173	0.1114	0.1066	.1014

They evaluated performance of the model over varying top number of pages recommended by the model. They used web server log of NASA data set for the period from July 1, 1995 to July 5, 1995 as training data set and July 6, 1995 as test data set. To compare the proposed SCFR system with hybrid model

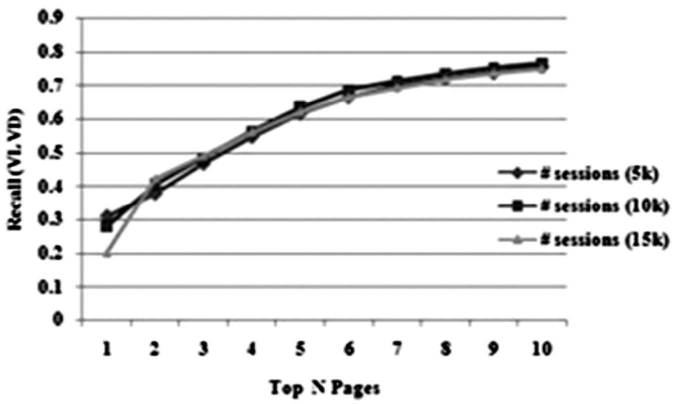

Figure 6: recall for 5k, 10k and 15k sessions with VLVD as distance measure

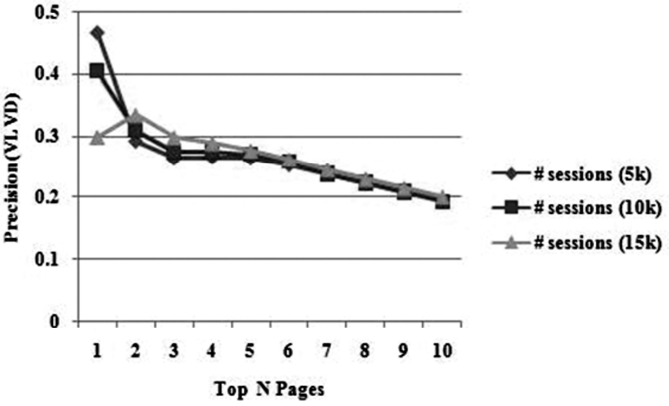

Figure 7: precision for 5k, 10k and 15k sessions with VLVD as distance measure

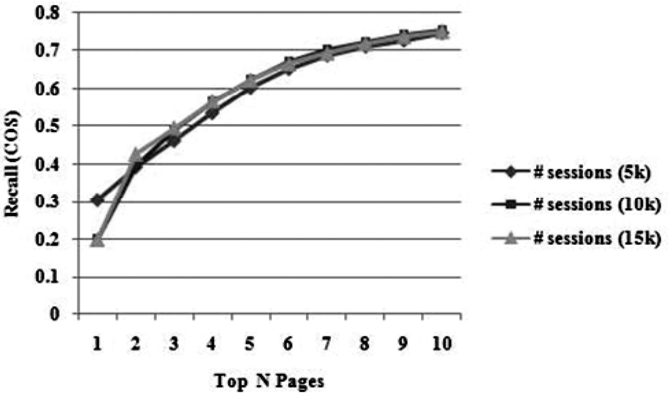

Figure 8: recall for 5k, 10k and 15k sessions with cosine similarity as distance measure

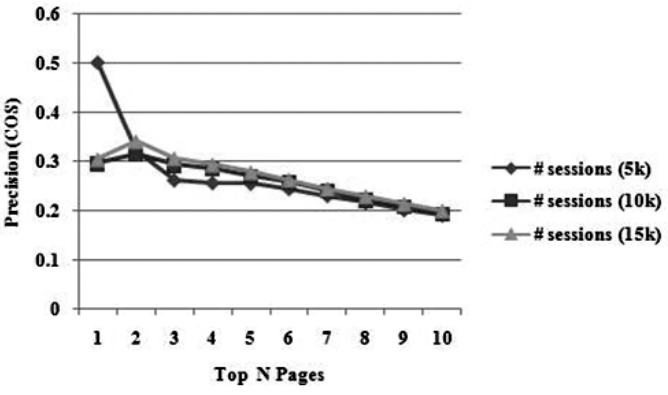

Figure 9: precision for 5k, 10k and 15k sessions with cosine similarity as distance measure

the same data set is used. The recall and precision of all these models is depicted in figure 10 and 11 respectively. The performance of the VLVD and cosine similarity measures used in SCFR are almost same and are much better than the various combination of models used in the hybrid model in terms of both recall and precision as can be seen from table 1 and 2. The graphs of figure 10 and 11 clearly show the goodness of the proposed SCFRS compare to other hybrid models in terms of both recall and precision. Also, for online recommendation the hybrid model may take more time because, it tries to cover the part of active session by various models in sequence and in the worst case all the four models may be required to give recommendations. In the proposed SCFR system the active session is assigned to the nearest cluster and recommendations are suggested based on the cluster to which active session is assigned. Hence proposed SCFR system is efficient compare to the hybrid click-stream based collaborative model.

4. CONCLUSION AND FUTURE WORK

An efficient recommender system based on collaborative filtering and user session clustering is proposed. The proposed SCFR system is evaluated using standard metrics recall and precision. The results obtained are compared with the hybrid recommender system that is based on markov model, sequential association rule, association rule and the default mode. The results clearly illustrate the goodness of proposed recommender system compare to the hybrid recommender system.

The future work would be to find the frequent pattern of the cluster and generate the recommended list based on this frequent pattern. The proposed SCFR system generates the recommendation list of pages by comparing the part of active session with the sessions of nearest cluster. If frequent pattern is determined for each cluster during off line phase, the time taken to compare active session with the session of cluster may be avoided.

REFERENCES

[1] J. B. Schafer, J. Konstan, J. Riedl, “Recommender systems in e-commerce,” First ACM Conference on Electronic Commerce, pp. 158–166, 1999.

[2] M. Adnan, M. Nagi, K. Kianmehr, R. Tahboub, M. Ridley, J. Rokne, “Promoting where, when and what? An analysis of web logs by integrating data mining and social network techniques to guide ecommerce business promotions,” J. Soc. Netw. Anal. Min., vol.1, pp. 173-185, 2011.

[3] C. Li, A. Datta, A. Sun, “Minig latent relations in peer-production environments: a case study with Wikipedia article similarity and controversy,” J. Soc. Netw. Anal. Min., vol. 2, pp. 265-278, 2012.

[4] T. Yan, M. Jacobsen, H. Garcis-Molina, D. Umeshwar, D, “From User Access Patterns to Dynamic Hypertext Linking,” The Fifth International World Wide Web Conference, pp. 1007-1014, May 1996.

[5] B. Mobasher, H. Dai, T. Luo, M. Nakagava, “Discovery and evaluation of aggregate usage profiles for web personalization,” J. Data Mining and Knowledge Discovery, vol. 6, no. 1, pp. 61-82, 2002.

[6] B. Mobasher, H. Dai, T. Luo, M. Nakagava, “Effective Personalization Based on Association Rule Discovery from Web Usage Data,” The third International workshop on Web Information and Data Management, pp. 9-15, 2001.

[7] R.Forsati, M. Meybodi, A. Rahbar, “An Efficient Algorithm for Web Recommendation Systems.” The IEE/ACM International conference on Computer Systems and Applications, pp. 579-586, 2009.

[8] P. Kazienko, “Mining indirect association rules for web recommendation,” Int. J. Appl. Math. Comput. Sci., vol. 19, no. 1, 165–186, 2009.

[9] B. S. Suryavanshi, N. Shin, S. P. Mudur, “Improving the effectiveness of model based recommender systems for highly sparse and noisy Web usage data,” The 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 618-621, 2005.

[10] A. Kumar, P. Tambidurai, “Collaborative web recommendation systems based on an effective fuzzy association rule mining algorithm.” Indian journal of computer Science and engineering, vol. 1, no. 3, pp. 184-191, 2010.

[11] W. Yong, L. Zhanhuai, Zhang Yang, “Mining sequential association-rule for improving Web document prediction,” The Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA), pp. 146-151, 2005.

[12] W. Ping, “Web page recommendation based on Markov Logic Network,” The third IEEE International Conference on Computer Science and Information Technology (ICCSIT), pp. 254-257, 2010.

[13] Y. Peng, G. Xiao, T. Lin, “Prediction of user’s behavior based on matrix clustering,” The Fifth International conference on Machine Learning and cybernatics, pp. 1343–1346, 2006.

[14] R. Katarya, O. P. Verma, “An effective web page recommender system with fuzzy c-mean clustering,” Multimedia Tools and Applications, pp. 1–16, 2016.

[15] X. Xianfen, B. Wang, “Web page recommendation via twofold clustering:considering user behavior and topic relation,” Neural Computing and Applications, pp. 1-9, 2016.

[16] Y. Wang, W. Dai, Y. Yuan, “Website browsing aid: A navigation graph-based recommendation system,” J. Decision Support Systems, vol.45, pp. 387-400, 2007.

[17] S. Salin, P. Senkul, “Using semantic information for web usage mining based recommendation,” The 24th International Symposium on Computer and Information Sciences, pp. 236–241, 2009.

[18] J. Li, O. R. Zaiance, “Combining usage, content and structure data to improve web site recommendation,” Lecture Notes in Computer Science: E- commerce and Web Technologies, vol.3182, pp. 305–315, 2004.

[19] M. Goksedef, S. Gündüz, “Combination of Web page recommender systems,” J. Expert Systems with Applications, vol. 37, pp.2911-2922, 2010.

[20] D. Kim, L. lm, N. Adam, V, Atluri, M. Bieber, Y. Tesha, “A Clickstream–Based Filtering Personalization Model: Towards A Better Performance,” The 6th annual ACM International workshop on web information and data management, pp. 88-95, 2004.

[21] J. Sobecki, “Web-based system user interface hybrid recommendation using ant colony metaphor,” Knowledge-Based Intelligent Information and Engineering Systems, LNAI (4694), pp. 1033–1040, 2007.

[22] N. Golovin, E. Rahm, “Reinforcement Learning Architecture for Web Recommendations,” The International Conference on Information Technology: coding and computing, pp. 398-402, 2004.

[23] D. H. Park, H. K. Kim, I.Y. Choi, J. Kim, “A literature review and classification of recommender systems research,” Expert Systems with Applications, vol. 39, no. 11, 10059–10072, 2012.

[24] R. M. Bell, Y. Koren, C. Volinsky, “Modeling Relationships at Multiple Scales to Improve Accuracy of Large Recommender Systems,” KDD’07, pp. 95-1042007.

[25] X. Su, T. M. Khoshgoftaar, “A survey of collaborative filtering techniques,” Advances in Artificial Intelligence, pp.1–19, 2009

[26] J. Bobadilla, F. Ortega, A. Hernando, Gutierrez, “Recommender systems survey,” Expert Systems with Applications, vol. 46, pp. 109-132, 2013.

[27] G. Poornalatha, S. R. Prakash, “Web User Session Clustering Using Modified K-means Algorithm,” The First International Conference on Advances in Computing and Communications, CCIS (191), pp. 243-252, 2011

[28] L. Wei, Z. Shu-hai, “A Hybrid Recommender System Combining Webpage Clustering with Web Usage Mining,” The International conference on Computational Intelligence and Software Engineering, pp. 1-4, 2009.