Multimed Tools Appl (2017) 76:21481–21496
DOI 10.1007/s11042-016-4078-7

An effective web page recommender system with fuzzy c-mean clustering

Rahul Katarya [1] **· Om Prakash Verma** [1]

Received: 22 February 2016 / Revised: 4 September 2016 / Accepted: 18 October 2016 /
Published online: 28 October 2016

Abstract With the exponential development of the number of users browsing the internet, an important factor that now the developer community is focussing on is the user experience. Recommender systems are the platforms that make personalized recommendations for a particular user by predicting the ratings for various items. Recommender systems majorly ignore the sequential information and rather focus on content information, but sequential information also provides much information about the behavior of the user. In this research work, we have presented a novel web-based recommender system which is based on sequential information of user's navigation on web pages. We received top-N clusters when Fuzzy C-mean (FCM) clustering is employed. We determined the similar users for the target user and also evaluated the weight for each web page. We have tried to solve that problem of recommender systems as we offered a system to forecast a user's next Web page visit. In our work, we proposed a system which generates recommendations to the users, by considering the sequential information that exists in their usage patterns of Web pages. We employed fuzzy clustering to give recommender system a sequential approach. We calculated weights for each page category considered in our system and predict top page recommendation for the target user. The real-world dataset of MNSBC is used in the experiments. The dataset consists of 5000 user entries with 6, entries per user. When we performed a comparison between the existing model with our proposed model, then it clearly showed that the accuracy of the proposed model is almost three times better than some existing systems. The accuracy of our proposed model is nearly 33 %.

Keywords Web recommender system · Clustering · Fuzzy c-mean · Accuracy

✉ Rahul Katarya
rahulkatarya@dtu.ac.in

[1] Department of Computer Science & Engineering, Delhi Technological University (Formerly Delhi College of Engineering), Shabad Daulatpur, Main, Main Bawana Road, Delhi 110042, India

1 Introduction

A collection of data in enormous amounts and affordable processing capabilities have motivated organizations to shift from the traditional world of mass production to the new world of customization. The development of e-commerce platforms has allowed companies to provide more options to the customers on a single integrated platform. The recommendation system is a decision support system which can offer the desirable information to the customers as per their needs [23, 30, 31, 61]. Recommender systems (RSs) are the platforms that make personalized recommendations for a particular user by predicting the ratings for various items [21, 28, 29, 32, 55]. The products can be recommended, by various factors such as top sellers on a site, demographics of the customer, social relation of users and past buying behavior of the client. RSs generate different recommendations to different ensemble users, thus providing customized web interface to the users. Currently, most of the e-commerce industries have enabled recommendation systems at the back end and offering personalized web recommendations to the users. Recommender systems have been developed by incorporating various disciplines such as machine learning, computational intelligence, and data mining techniques and especially heuristics and finding association patterns among the items [11, 57, 63]. Examples of popular recommender systems include Amazon.com for books, CDs and various other products, IMDB for movies and Jester system for jokes. Advancement in techniques of different domains like machine learning, pattern recognition, and statistics have made it possible to perform mining and unfold exciting as well as unknown patterns from the big data. Web data exists in several formats like URL visits, web page content, incoming and outgoing hyperlinks to the Web page. By what type of data is being analysed or mined, so web mining was broadly classified into web content, web usage and web structure mining [38, 39, 52]. Web structure mining works on the internet's hyperlink structure. Web usage mining develops innovative, implicit and valuable patterns from the usage data of users. Web page recommendation systems have quite a persistent approach to analyzing users' behavior over the internet and produce recommendations as per their preferences. It supports organizations in intelligent decision making on their customers' needs by automating the recommendations as per their preferences. While building recommender systems, the sequential characteristic of the user session is ignored. A sequential aspect of web user session has been considered by probabilistic models like Markov model while designing a web recommendation system [41]. However, the problem with the probabilistic model is that swapping probability between web categories should be known a priori, and may involve the knowledge and experience of a domain expert. In our work we have proposed a system which produces recommendations to the users, by considering the sequential information that exists in their usage patterns of web pages. We employed fuzzy c-mean clustering technique to generate soft clusters among our users and picked up top clusters for processing; that depend on the average value of cluster centers. We calculated weights for each page category that were considered in our system and predicted top page recommendation for the target user. We found out accuracy of our prediction with existing MSNBC dataset. We conducted several experiments on a large dataset of MSNBC that consisted of 5000 users with six-page visit data and predicted the seventh page. The experiment results show that our proposed model can better utilize user's social information and outperforms the modern algorithms regarding accuracy. The leading contributions of our research work are summarized as:

- We have presented a recommender system which generates recommendations for the users, by considering the sequential information that exists in their usage patterns of web pages.
- Our proposed system considered both the content as well as the sequence of a visit during clustering to form groups of users.
- We employed fuzzy c-mean clustering techniques to generate soft clusters among our users.
- We selected top clusters for processing which depended on the average value of cluster centers.
- We calculated weights of each considered page category and predicted top page recommendation for the target user.
- We found the prediction accuracy with existing MSNBC dataset and conducted several experiments on a large dataset of MSNBC which consists of 5000 users with six-page visit data and predicted the users' seventh-page visit.
- Our proposed system was compared with state of art methods such as WRS model [41], WebPUM [24], Modified K-means [45], Multivariate [9], SemSig [20], Weighted K-means [53], Binary data clustering [16], Neuro-Fuzzy [50], XFCM [54] and C-means & COG [33].
- The experiment results showed that our proposed system performed with better results, and it can be utilized for acquiring user's social information over Web navigation based recommender system.

The remainder of this article is organized as follows: Section 2 gives a review of the background knowledge of our work. Section 3 presents our novel web page recommendation model. Section 4 describes the experiments and the performance of our proposed system, and finally, Section 5 concludes our work.

2 Related work

Various studies have been performed in improving the efficiency of recommender systems for better recommendations to the user. Recommendation systems are designed using different methods by including k-NN, decision tree, clustering, regression, heuristic approaches, neural networks and association rule mining [2, 13, 64]. Based on the kind of techniques used, recommendation systems can be categorized as content-based and collaborative based systems. The content-based approach has originated from the information retrieval and information filtering domain [5, 12]. Content-based recommender systems make recommendations based on users' past likings. The rating for any item for any individual is computed based on ratings of similar items given by the person, and primary objective is to study a function that forecasts that in which class a document fits (i.e., either adored or not liked) [6, 36]. Collaborative systems are different from content-based systems in the sense that they first found related users for target users and then make recommendations based on preferences of same users [67]. In this methodology, recommendations are made by finding correlations among the users. The key objective of the collaborative filtering is to discover rating of the items, which are not seen by the current user and recommended the ratings to similar users. Collaborative based recommender systems can be more classified into two classes, memory based (heuristic based) and model based collaborative systems [4, 15]. Memory based systems calculate the similarity among users based on users' ratings. The algorithms of memory based systems are heuristics that make recommendations based on a whole collection of items pre-rated by the users. Model-based collaborative recommender systems generate the descriptive model of the system which are based on the users' preferences, numerous data mining, and machine learning techniques. The several methods are used such as Bayesian models, clustering models, and latent semantic models. There are various other probabilistic modeling techniques which are used for building recommender systems. Ravi et al. [46] presented a comprehensive, state-of-the-art review of the research work done on opinion mining and also examined in seven broad dimensions viz. Subjectivity classification, sentiment classification, review usefulness measurement, lexicon creation, opinion word, product aspect extraction, opinion spam detection and various applications of opinion mining. Web-page recommendation plays a significant role in smart web systems. Beneficial knowledge discovery from web usage data and adequate knowledge representation for expert web page recommendations are crucial and challenging tasks. Clustering is an essential phase in the web mining, as the data on the web is increasing in a huge way. A multitask spectral clustering model was proposed to handle the several data clustering challenges and then suggested two types of correlations such as inter-task clustering correlation, which discussed the relations among different clustering tasks and intra-task learning correlation, which permitted the procedures of learning cluster labels [60]. In the data clustering research, a clustering algorithm was proposed (discriminative nonnegative spectral clustering) that explicitly imposed an additional nonnegative constraint on the cluster indicator matrix to pursue a more interpretable resolution [59]. Nguyen et al. [43] proposed a technique to provide proficiently enhanced web page recommendation through semantic-enhancement by mixing the domain and web usage knowledge of a website. Two new models were offered to represent the domain knowledge. The first model used an ontology to represent the domain knowledge. The second model used automatically generated semantic network to represent domain terms, web-pages, and the relations between them. A solution for a cross-site cold-start product recommendation was presented, which focused on recommending goods from e-commerce portals and websites to users at social networking sites in cold-start states [66]. E-Commerce businesses have accepted web personalized procedures broadly in the form of recommender systems for prompting user behavior for customer preservation. Some studies were performed by taking three significant assistances to the body of knowledge in information systems. The first study was based on web personalization, in the second study a Utaut2 model was presented and in the third study, web personalization was dedicated on e-commerce domain [34]. An effective recommender system was proposed by handling the session data of users and presented an algorithm to divide the binary session data into a static number of groups and applied the partitioned sessions to make recommendations

[16]. Location-based social networks (LBSN) enable users to check-in at points of interests (POIs), share the information with other users on the network, and receive recommendations about new and interesting POIs in their vicinity. Schmachtenberg et al. [49] showed an approach for enhancing POI recommendations for a location-based social network by enriching it with semantic data. Social networks are playing a significant role in real-time information exchange and online human interaction such as Facebook, Twitter, LinkedIn, Amazon and Sina Weibo. The enormous volume of live information in social media stream-shave directed to the extraordinary requirement for tracking technologies. In addressing this issue, a multi-faceted brand tracking process was proposed to gather representative data from large-scale social media content [17]. In their technique, authors took benefits of the heterogeneous data of social media content. Social websites and microblogs posted on social network comprise an abundant percentage of images. Henceforth how to identify logos in images from the social network is of high value. To address this issue, a learning-based logo detection technique was suggested with social network information supporter, in which dense histogram type feature was also proposed to categorize logo and non-logo image patches [56]. Event discovery in social media has become an essential task. The foremost challenges in event detection lie in the features of social media data, such as little, informal, mixed and live. By considering all these factors, an event detection method was offered, which produced an intermediate semantic entity, named microblog clique (MC), to discover the highly correlated information among the noisy and small microblogs [18]. The Markov model is complex probabilistic model widely used for modeling sequential events. Several hybrid recommendation systems have also been developed using multiple techniques including a sequential pattern analysis. Maximum of the work related to the design of a guidance and recommended a system with subsequent data used the Markov model [25, 26]. Sequential and association pattern mining procedures have been established to find the sequential arrangements in the data. Our proposed model considers both the content as well as the sequence of a visit during clustering to form groups of users. Clustering is the approach of data mining which is adopted by various domains of web and information systems. A fuzzy diagonal co-clustering algorithm with K-means was proposed which were evaluated on synthetic and real datasets [35]. Web page recommender system has drawn attention to many researchers and industrialists as it offers attractive schemes and after regular intervals of time. So to suggest the suitable web page from the large collections of pages, a Tclus algorithm based system was presented in which authors combined the features of density based and k-means clustering algorithm by collecting the data from the website of around five months [58]. Whenever we talk about the internet based recommender system, one aspect which we will always need that is decent value of web service, but most of the web recommendations are effected with the sparsity and scalability problems. To reduce these problems, a ClucF method was offered which utilized the users and services clusters [62]. In the similar work, authors tried to the refine web recommender system in which they utilized the k-means clustering technique and compared with modified knockout refinement algorithm (MKRA) [14]. A neuro-fuzzy based system was offered in which authors tried to get patterns and browsing behavior from the web log data [50]. The neuro-fuzzy system was mixed with the neural networks and the fuzzy set theory. The browsing

behavior analysis for websites is necessary for detecting suspicious activities for online users. A news article recommender system was designed in which collaborative filtering and k-means clustering approaches were employed [8]. Authors used the WordNet-enabled k-means algorithm, which discovered the usage of word hypernyms mined from the WordNet database. In another research work for web mining, a movie recommender system was presented in which they handled the issues such as sparsity, scalability, new users and cold-start problems and they united the web mining approaches and domain specific ontologies [42]. A study was presented in which authors advised the role of agent technology in web usage mining with homomorphic encryption for e-commerce environment and they also highlighted the user's privacy and agents [51]. In the related study, the importance of the semantic web in the knowledge discovery and data mining fields were analyzed [47]. The research was accomplished in which growth of blended learning and success of education with ICT tools were taken into consideration. The blended model was mixed with a learning management system (LMS), a Web 2.0 tools and the E-learning activities recommender system (ELARS) [22]. Some authors presented a methodology to discuss the web framework which was based on observation system for earth program. Their operational tactic was composed of five technologically neutral stages to form executable business practices [48]. A personalized social document representation (PerSaDor), was presented which was based on the social media in which authors presented two ranking functions for documents which were established on their textual content and their PerSaDoR [7]. A semantic methodology was offered in which ontology was used as an arbitrated plan for the tracking data from web source's semantics. Semantic mappings among the scheme of origin and the ontology were used to transform the original data to RDF [19]. A system named as TANGO was offered which role was to study rules to abstract and fetch the information from semi-structured web pages with high precision and recall for enterprise systems. TANGO was based on an open directory of features that was characterized by resources of HTML, DOM, CSS and user-defined features [27]. A method was presented to anonymize the query logs which employed the k- anonymity and de-anonymization tools and finally regulated the possible privacy problems [44]. A study was dedicated to the dynamic features of the Kohonen procedure, which was responsible for transporting proficient incremental procedure [37]. Big data are generated from the web transactions and interactions. Unsupervised learning approaches are cooperative in separating heterogeneity of treatment effects for such actions. By taking these issues under consideration authors designed multiple imputations (MI) based fuzzy clustering and suggested a new MI-based Visualization-aided validation index (MIVOOS) in contrast to fuzzy clustering validation indexes, XB and VOS [65]. Researchers presented a method which designed the frequent page sets from weblog in the dynamic environment. They proposed T + weight tree organization and concluded by experiments that it was more competent than weighted tree method with respect to time and space [40].

3 Proposed work

Clustering of data is the process of dividing data into classes or groups so that items in the matching group are as similar as possible, and items in diverse classes are as dissimilar as possible. Based on the nature of data and the determination for which clustering is being used, several actions of similarity may be used to place various items into classes, where the similarity measure controls how the clusters are formed. The criteria for identifying similar users is the browsing history of users. The users who have opened nearly same category of pages as our target user is likely to be in the same cluster to our target user. A pattern recognition based recommender system involves two stages; the first step is clustering followed by classification task. At the first stage, the system is provided with sufficient learning so that the classification accurateness of the system is quite significant or at the desired level. After the system learns, it generates a set of recommendations with good rankings. The new user is classified among clusters that we have taken with the help of FCM algorithm. As soon as a new user enters the system, we need to apply FCM over the whole dataset again so as to allocate each user a cluster number after that recommendation is made.

Figure 1 gives the overall flow of the model and the order of occurrence of events. First of all, in the given MSNBC dataset we employed the fuzzy c-mean clustering approach. By applying the method, we get the top-N clusters that are further used for identifying the similar users for a target user. Then we calculate the weight for each page category and finally determine the accuracy of the system.

3.1 Fuzzy c-mean clustering

In our system, we first designed clusters to obtain knowledge about web users and the classification method was used later for increase the knowledge fitness and produced recommendations. A web user may have multiple interests for which he needs to be put into multiple clusters. Hence, we have used fuzzy c-mean technique [1, 3, 10]. In fuzzy clustering or soft clustering, data elements can go to more than one cluster, and linked to each item is a set of membership levels. These specify the strength of the relationship between that data element and a particular cluster. Fuzzy clustering is a procedure of assigning these membership levels, and then using them to assign data elements to one or more clusters. The FCM algorithm efforts to partition a finite collection of n elements $X = \{X_{1,\ldots,}X_{nj}\}$ into a group of c fuzzy clusters on some given standard. Given a fixed set of data, the algorithm takes a list of c cluster centres $C = \{c_{1,\ldots,}c_c\}$.Moreover, a partition matrix W = $w_{i,j} \in$ [0, 1], i = 1,…,n, j = 1,…c, where each element w_{ij} given by Eq. 1 says the degree to which element X_i fits to cluster c_j. Like the K-means clustering, the FCM purposes to minimize an objective function.

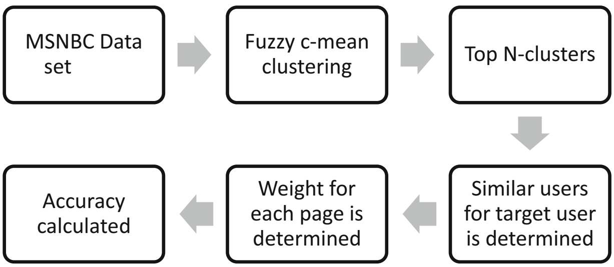


Fig. 1 Block diagram of our proposed model

$$c^{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \| \left(X_i - c_j \right) \|^2 \tag{1}$$

Where,

$$w_{ij}^{m} = {}^{1}\!/\!\sum_{k=1}^{c} \left[\frac{\| \left(X_i - c_j \right) \|}{\| \left(X_i - c_k \right) \|} \right] \frac{2}{m-1}$$

This alters from the k-means objective function by the addition of the membership values w_{ij} moreover, the Fusilier $m \in R$, with $m \geq 1$. The Fusilier m regulates the level of cluster fuzziness. A large m consequences in smaller memberships w_{ij} and henceforth, fuzzier clusters. In the limit $m = 1$, the memberships w_{ij} converge to 0 or 1, that suggests a crisp splitting. In the absence of experimentation or domain knowledge, m is commonly set to 2.

3.2 Selection on top 'N′ clusters

After applying fuzzy c-mean clustering, we get cluster center values for each cluster upon which processing is done instead of entire data of users present in the cluster. We are required to find out the cluster number in which each user occurs. Since we are dealing with soft clusters, we may get a number of cluster numbers for each user. That is acceptable. We are taking top 'n' clusters in our proposed work depending on the average value of cluster center values. It comes from the fact that not all but significant clusters will have a remarkable impact on recommendations.

3.3 Finding users similar to target user

We now choose our target user for whom we wish to make recommendations. After choosing target user we now need to find all users similar to target user as they would have a direct impact on user's recommendations. To find related users, we used partition matrix 'U,' that is obtained from fuzzy c-mean clustering algorithm. The similar users are found for the target user using Fuzzy c-means clustering algorithm which generates soft clusters i.e. A user can be present in more than one cluster. After that, Top N-clusters are produced for the user. We got similar users for target user which we generated through resultant matrix MAT, and which also contains all similar user entries. When a new user is encountered with having some browsing history, our system is trained to handle such cases as well. The system will first find out a similar cluster of the new user with the already existing users and after that recommendation is made for the new user as well.

3.4 Weight calculation for each page category

Since we have obtained a set of similar users, we are left with the calculation of weight for each page category. Higher the weight, higher is the probability of the page to be recommended to the user. In our proposed model we are having 17-page categories defined with ranking from 1 to 17. The data set has seventeen categories on the front page, weather, health, living, business, news, tech, local, opinion, on-air, misc., sports, summary, bbs(bulletin board service), travel, MSN-news, and MSN-sports. For the purpose of weight calculation, we maintained three matrices namely 'X,' 'Z' and 'G.'

- Matrix X contains a count of a number of times the target user has opened the specific page.
- Matrix Z contains a count of a number of times all the similar users of target user have opened the specific page.
- Matrix G contains whether a page has been opened by the target user at all or not.

Taking into consideration all these matrix values it can be understood that matrix X and matrix Z favours recommendation of a page to the user. Corresponding mathametical expressions are as follows:

$$\begin{aligned}&\text{if (msndata (user, i) == k)}\\&\text{x(k) = x(k) + 1;}\end{aligned} \tag{2}$$

$$\begin{aligned}&\text{if(mat (i, j) == k)}\\&\text{z(k) = z(k) + 1;}\\&\text{if (mat (i, j) == k)}\\&\text{if (g(k) == 0)}\end{aligned} \tag{3}$$

$$\text{g(k) = g(k) + 1;} \tag{4}$$

Where k stands for page category from 1 to 17, matrix mat contains all similar users entries of the target user, and msndata refers to dataset taken for analysis. At last, when all the above three matrix values have been calculated the final weight of each page category can be computed using the mathematical expression stated below:

$$\text{Weight (i)} = (\text{X[i]} + \text{Z[i]})/(\text{size}-\text{G[i]}) \tag{5}$$

Where i stands for page category and values used can be obtained using Eqs. 2, 3 and 4.

"size" is a number of users in the cluster of the target user obtained after applying FCM algorithm over the dataset.

3.5 Calculation of accuracy

Accuracy is the proportion of the amount of correct recommendations to the volume of whole recommendations.

Accuracy = (number of accurate recommendations) / (number of whole recommendations)

In this case, we recommend the next likely visit of the user, which is matched with the authentic next step of the user (testing). If the predicted following step is the same of the actual next state, then the event is termed as hit, else it is termed as a miss. Hence, accuracy is our case will be given as:

$$\text{Accuracy} = (\text{number of hits})/(\text{number of hits} + \text{number of miss}) \tag{6}$$

Considering the confusion matrix accuracy of the recommendation system can be defined as the ratio of significant retrieved elements to all retrieved, non-retrieved, relevant and not relevant elements. The corresponding algorithm used in the proposed work has been stated in Fig. 2.

We presented a web-based recommender system which generates recommendations of web users, by considering the sequential information that exists in their usage patterns of web pages

Input: 5000 X 6 MSNBC dataset, where a number of users are 5000, and each column contains page category number varying from 1 to 17.

1. Apply fuzzy c-mean clustering on MSNBC dataset to obtain soft clusters.
2. Select top 'N' clusters among all clusters depending on the average value of cluster centers obtained from equation 1.
3. Select the target user and find all users similar to the target user and generate matrix MAT.
4. Obtain values of matrix 'X' using the equation 2.
5. Similarly, obtain values of matrix 'Z' and 'G' using equations 3 and 4.
6. Finally, calculate a weight for each page category using equation 5, and determine values of matrix page which contain predicted page for each user.
7. Calculate accuracy of the predicted result using Equation 6.

Output: Similar users for the target user. It gives two matrices:
Center Matrix: It contains average value of columns
Partition matrix: it contains a contribution of each user in each cluster.

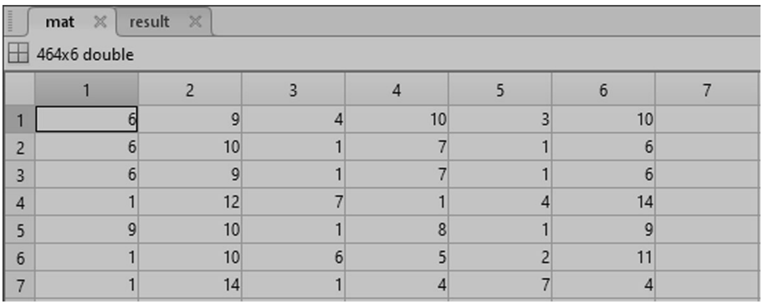
Fig. 2 A high-level description of the pseudocode of the proposed model

in MSNBC dataset. The proposed system considered both the content as well as the sequence of a visit during clustering to form groups of users. The fuzzy c-mean clustering procedure was employed to produce soft clusters among users and nominated limited top clusters for processing which depended on the average value of cluster centers. We computed weights of each considered web page category and predicted top page recommendation for the target user and calculated the accuracy of prediction with existing MSNBC dataset. We conducted experiments on a large dataset of MSNBC which contained 5000 users with six page visited data and predict the seventh page which has to be visited. Our proposed system was compared with state of art methods such as WRS model [41], WebPUM [24], Modified K-means [45], Multivariate [9], SemSig [20], Weighted K-means [53], Binary data clustering [16] and Neuro-Fuzzy [50]. The experiment results displayed that our proposed model performed with better results, and it can be applied for acquiring user's social information, pattern, personality and behavior over web navigation in a web-based recommender system.

4 Results and discussions

4.1 Dataset

In this paper, a real-world dataset of MNSBC is used in the experiments. The dataset consists of 5000 user entries with 6 entries per user. The dataset is distributed into two parts as follows: training data = 60 % and testing data = 40 %. The number of clusters used = 100 and number


mat | result

464x6 double

	1	2	3	4	5	6	7
1	6	9	4	10	3	10	
2	6	10	1	7	1	6	
3	6	9	1	7	1	6	
4	1	12	7	1	4	14	
5	9	10	1	8	1	9	
6	1	10	6	5	2	11	
7	1	14	1	4	7	4	

Fig. 3 Recommendation without sequential information without seventh users

of users in dataset = 5000. Entries for each user are a sequence of numbers from 1 to 17, representing browsed page history for the user. We applied our proposed model on 6 entries per user to predict the 7th page, which is to be browsed by the user. The results obtained are then checked with 7th column entry of the MSNBC dataset and accuracy is calculated by comparing values of both. For the experimental setup, we used a machine having a fourth generation Intel Core I5 processor coupled with 6GB RAM.

4.2 Results

Clearly, it can be observed from the experimentation results that our proposed model was three times efficient than WRS model [41] and also delivered proficient results when compared to other existing systems.

In Fig. 3, an input matrix for first seven users is presented without the sequential information. We need to find a 7th entry for the user. The data contains page category numbers that user browsed in the past and with the help of those, we will predict 7th-page category that user is more likely to open. In Fig. 4, output matrix for first seven users is presented with sequential information.

Table 1 shows the comparison between various existing systems. The accuracy of our proposed model is far better than the earlier proposed systems. The accuracy of our proposed model is nearly 33 % that is 1 out of every 3 predictions is correct while the accuracy of the previous system was 1 hit and 9 miss. Time should be overlooked as the proposed model provides around three times accuracy than existing method such as WRS model. We have considered a range of 20 to 32 clusters. Table 1 and Fig. 5 show the comparison among different systems with our proposed model. It clearly indicates that the accuracy of the proposed model is much better than already existing systems such as WRS model, WebPUM, Modified K-means, Multivariate, SemSig, Weighted K-means, Binary data clustering, Neuro-Fuzzy, XFCM and C-means & COG.

For the M = 20, we compared our model with other existing methods and found that WebPUM has accuracy as 22.322 as it was based on the graph partition algorithms to detect patterns of web pages. Modified K-means acquired 27.642 % of accuracy, as it was based on the fact that which computed the distances between the similar sessions of web access. In the multivariate method, the major focus was on to analyze the properties and temporal behavior of content and material of the internet pages; the accuracy percentage was retrieved as 30.274. The SemSig was based on the fact of entity linking on web pages, for M = 20 this system delivered the accuracy percentage as 29.543. We also evaluated a system which was based on the Weighted K-means, in which quality of web page recommendation was enhanced with the utilization of a combination of user profiles and mean square residue. The Weighted K-means system gave accuracy percentage as 28.285. We also evaluated a system which was based on the binary data clustering which handled the binary session and formed the fixed numbers of

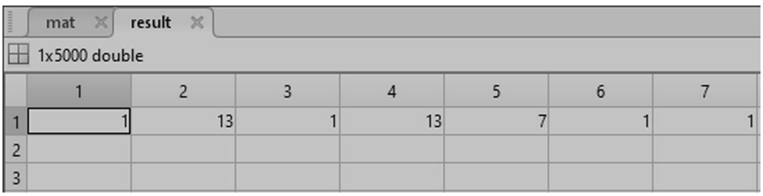


Fig. 4 Recommendation with sequential information of with the seventh user

Table 1 The comparisons with the various systems on a different number of clusters

Methods Accuracy (%) / No. of clusters(M)	M = 20	M = 21	M = 22	M = 32
WRS Model [41]	10.260	10.821	11.295	11.062
WebPUM [24]	22.322	22.872	23.364	23.964
Modified K-means [45]	27.642	27.537	27.923	28.769
Multivariate [9]	30.274	30.862	31.883	31.896
SemSig [20]	29.543	29.784	30.324	30.867
Weighted K-means [53]	28.285	28.384	28.968	29.278
Binary data clustering [16]	28.926	29.272	29.475	29.847
Neuro-Fuzzy [50]	31.537	31.496	31.865	32.245
XFCM [54]	32.672	32.837	32.973	33.154
C-means & COG [33]	31.453	31.839	32.454	32.892
Proposed Model	32.760	32.920	33.160	33.100

clusters by utilizing clustering algorithm. Binary data clustering based system provided the accuracy percentage as 28.926. A Neuro-Fuzzy based system was applied on the web recommender system which resulted in the accuracy percentage as 31.537. The exponential fuzzy clustering (XFCM) was based on clustering objective with exponential mathematical equation and delivered the accuracy percentage as 32.672 at M = 20. C-means & COG based system was evaluated and delivered the accuracy percentage as 31.453 as this system was based on the mixture of Fuzzy c-mean and center of gravity. In Table 2, we have also compared the performance of various existing systems with respect to time for the MSNBC dataset.

5 Conclusion and future work

Web recommendation has attracted much attention from both academic and industry. In this article, we have suggested a novel model to expect a user's succeeding web page visit. The experiments have been accomplished on the MSNBC dataset which is web navigation dataset.

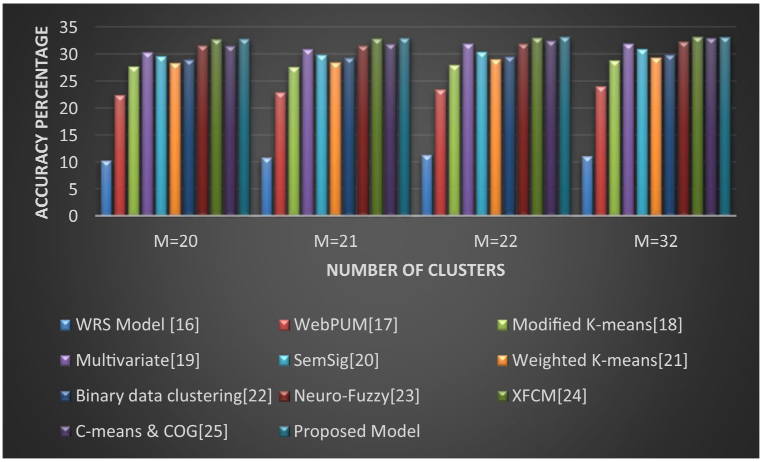


Fig. 5 Comparison with various existing systems with proposed model

Table 2 Comparision of speed for different systems for MSNBC dataset

Method/System	Time (in seconds)
WRS Model [41]	477
WebPUM [24]	498
Modified K-means [45]	277
Multivariate [9]	456
SemSig [20]	363
Weighted K-means [53]	282
Binary data clustering [16]	389
Neuro-Fuzzy [50]	410
XFCM [54]	428
C-means & COG [33]	415
Proposed Model	402

The outcomes have demonstrated the practicability of our proposed model. The subsequent visit of a user can provide useful evidence about their likes and perception. This statistics can be explored and utilized for projecting the desired category or product to the user. The probability of purchase for e-commerce companies can be enhanced, which in turn will growth the expected revenue for e-commerce organizations. The best usage of proposed system in electronic commerce (e-commerce) can be understood as follows: suppose a person wants to buy a television and search on Google but he has not checked it over any of the e-commerce websites. In that case with the help of proposed system e-commerce websites can come to know about what user is interested in buying and what should he be recommended even when the user is not explicitly looking for the product on their website also. For new websites, the proposed system is of great significance, as being new websites people will not be more likely to use them due to trust and reliability issues. However with the suggested system website owners can track that what user is going to purchase and can provide related recommendations to the user without their explicitly searching over their website. Our future work will be focused on inclusion of privacy, trust and social networks with the utilization of hybrid intelligent systems.

References

1. Baraldi A, Blonda P (1999) A survey of fuzzy clustering algorithms for pattern recognition. I. IEEE Trans Syst Man Cybern B Cybern 29:778–785
2. Barragáns-Martínez B, Costa-Montenegro E, Juncal-Martínez J (2015) Developing a recommender system in a consumer electronic device. Expert Syst Appl 42:4216–4228
3. Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy *c*-means clustering algorithm. Comput Geosci 10: 191–203
4. Bilge A, Gunes I, Polat H (2014) Robustness analysis of privacy-preserving model-based recommendation schemes. Expert Syst Appl 41:3671–3681
5. Bobadilla J, Ortega F, Hernando A, Gutiérrez A (2013) Recommender systems survey. Knowledge-Based Syst 46:109–132
6. Boratto L, Carta S (2014) The rating prediction task in a group recommender system that automatically detects groups: architectures, algorithms, and performance evaluation. J Intell Inf Syst:1–25

7. Bouadjenek M, Hacid H, Bouzeghoub M, Vakali A (2016) PerSaDoR: Personalized social document representation for improving web search. Inf Sci (Ny) 369:614–633. doi:10.1016/j.ins.2016.07.046
8. Bouras C, Tsogkas V (2014) Improving news articles recommendations via user clustering. Int J Mach Learn Cybern 1–15. doi:10.1007/s13042-014-0316-3
9. Calzarossa MC, Pavia FI, Tessera D (2014) Multivariate analysis of web content changes. In: 2014 IEEE/ ACS 11th Int. Conf. Comput. Syst. Appl. IEEE, pp 699–706
10. Cannon RL, Dave JV, Bezdek JC (1986) Efficient implementation of the fuzzy c-means clustering algorithms. IEEE Trans Pattern Anal Mach Intell 8:248–255
11. Cao J, Li Q, Ji Y et al (2016) Detection of forwarding-based malicious URLs in online social networks. Int J Parallel Prog 44:163–180
12. Cobo MJ, Martínez MA, Gutiérrez-Salcedo M, et al. (2015) 25years at knowledge-based systems: a bibliometric analysis. Knowledge-Based Syst
13. Conforti R, de Leoni M, La Rosa M et al (2015) A recommendation system for predicting risks across multiple business process instances. Decis Support Syst 69:1–19
14. Dixit VS, Bhatia SK (2015) Refinement and evaluation of web session cluster quality. Int J Syst Assur Eng Manag 6:373–389
15. Dooms S, Audenaert P, Fostier J et al (2014) In-memory, distributed content-based recommender system. J Intell Inf Syst 42:645–669
16. Forsati R, Moayedikia A, Shamsfard M (2015) An effective web page recommender using binary data clustering. Inf Retr J 18:167–214
17. Gao Y, Wang F, Luan H, Chua T-S (2014) Brand data gathering from live social media streams. Icmr:169–176
18. Gao Y, Zhao S, Yang Y, Chua T (2015) Multimedia social event detection in microblog. In: 21st Int. Conf. MMM 2015, Sydney, Aust. January 5-7, 2015. pp 269–281
19. García MDMR, García-Nieto J, Aldana-Montes JF (2016) An ontology-based data integration approach for web analytics in e-commerce. Expert Syst Appl 63:20–34
20. Guo Z (2014) Entity linking with a unified semantic representation. In: Int. World Wide Web Conf. Committee. ACM, pp 1305–1309
21. Hasija H, Katarya R (2014) Secure code assignment to alphabets using modified ant colony optimization along with compression. Proc 2014 Int Conf Adv Comput Commun informatics, ICACCI 2014:175–181
22. Hoic-Bozic N, Holenko Dlab M, Mornar V (2015) Recommedner System and Web 2.0 Tools to Enhance Blended Learning Model. IEEE Trans Educ in press:39–44
23. Hu X, Zeng A, Shang M-S (2016) Recommendation in evolving online networks. Eur Phys J B 89:46
24. Jalali M, Mustapha N, Sulaiman MN, Mamat A (2010) WebPUM: a web-based recommendation system to predict user future movements. Expert Syst Appl 37:6201–6212
25. Javari A, Jalili M (2015) A probabilistic model to resolve diversity–accuracy challenge of recommendation systems. Knowl Inf Syst 44:609–627. doi:10.1007/s10115-014-0779-2
26. Ji K, Sun R, Shu W, Li X (2015) Next-song recommendation with temporal dynamics. Knowledge-Based Syst 88:134–143
27. Jiménez P, Corchuelo R (2016) On learning web information extraction rules with TANGO. Inf Syst 62:74–103
28. Katarya R, Jain I, Hasija H (2014) An interactive interface for instilling trust and providing diverse recommendations. In: IEEE Int. Conf. Comput. Commun. Technol. ICCCT-2014. pp 17–22. doi:10.1109/ICCCT.2014.7001463
29. Katarya R, Verma OP (2015) Restaurant recommender system based on psychographic and demographic factors in mobile environment. In: IEEE Int. Conf. Green Comput. Internet Things 2015. pp 907–912. doi:10.1109/ICGCIoT.2015.7380592
30. Katarya R, Verma OP (2016a) A collaborative recommender system enhanced with particle swarm optimization technique. Multimed Tools Appl 75:1–15
31. Katarya R, Verma OP (2016b) Recent developments in affective recommender systems. Phys A Stat Mech its Appl 461:182–190
32. Katarya R, Verma OP, Jain I (2013) User behaviour analysis in context-aware recommender system using hybrid filtering approach. Proc - 4th IEEE Int Conf Comput Commun Technol ICCCT 2013:222–227
33. Koohi H, Kiani K (2016) User based collaborative filtering using fuzzy C-means. Measurement 91:134–139
34. Krishnaraju V, Mathew SK, Sugumaran V (2015) Web personalization for user acceptance of technology: an empirical investigation of E-government services. Inf Syst Front 18:579–595
35. Laclau C, Nadif M (2016) Hard and fuzzy diagonal co-clustering for document-term partitioning. Neurocomputing 193:133–147
36. Li B, Zhu X, Li R, Zhang C (2014) Rating knowledge sharing in cross-domain collaborative filtering. IEEE Trans Cybern 45:1–15
37. Liu D, Zhang Z, Guo X (2016) Web mining based on one-dimensional Kohonen's algorithm: analysis of social media websites. Neural Comput Appl 1–5. doi:10.1007/s00521-016-2410-9

38. Lorentzen DG (2014) Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. Scientometrics 99:409–445
39. Lotfy HMS, Khamis SMS, Aboghazalah MM (2015) Multi-agents and learning: implications for WebUsage mining. J Adv Res 7:285–295. doi:10.1016/j.jare.2015.06.005
40. Malarvizhi SP, Sathiyabhama B (2016) Frequent pagesets from web log by enhanced weighted association rule mining. Cluster Comput 19:269–277
41. Mishra R, Kumar P, Bhasker B (2015) A web recommendation system considering sequential information. Decis Support Syst 75:1–10
42. Moreno MN, Segrera S, López VF et al (2015) Web mining based framework for solving usual problems in recommender systems. A case study for movies' Recommendation. Neurocomputing 176:72–80
43. Nguyen TTS, Lu HY, Lu J (2014) Web-page recommendation based on web usage and domain knowledge. IEEE Trans Knowl Data Eng 26:2574–2587
44. Pàmies-Estrems D, Castellà-Roca J, Viejo A (2016) Working at the web search engine side to generate privacy-preserving user profiles. Expert Syst Appl 64:523–535
45. Poornalatha G, Raghavendra PS (2011) Web user session clustering using modified K-means algorithm. In: Adv. Comput. Commun. pp 243–252. doi:10.1007/978-3-642-22714-1_26
46. Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Syst 89:14–46. doi:10.1016/j.knosys.2015.06.015
47. Ristoski P, Paulheim H (2016) Semantic web in data mining and knowledge discovery: a comprehensive survey. Web Semant Sci Serv Agents World Wide Web 36:1–22
48. Santoro M, Nativi S, Mazzetti P (2016) Contributing to the GEO model web implementation: a brokering service for business processes. Environ Model Softw 84:18–34
49. Schmachtenberg M, Strufe T, Paulheim H (2014) Enhancing a location-based recommendation system by enrichment with structured data from the web. In: Proc. 4th Int. Conf. Web Intell. Min. Semant. - WIMS '14. pp 1–12. doi:10.1145/2611040.2611080
50. Shivaprasad G, Reddy NVS, Acharya UD, Aithal PK (2015) Neuro-fuzzy based hybrid model for web usage mining. Procedia Comput Sci 54:327–334
51. Sobitha Ahila S, Shunmuganathan KL (2016) Role of agent Technology in Web Usage Mining: homomorphic encryption based recommendation for E-commerce applications. Wirel Pers Commun 87:499–512
52. Thanh T, Nguyen S, Lu HY, Lu J (2014) Web-page recommendation based on web usage and domain knowledge. IEEE Trans Knowl Data Eng 26:2574–2587
53. Thiyagarajan R, Thangavel K, Rathipriya R (2014) Recommendation of web pages using weighted K-means clustering. Int J Comput Appl 86:44–48
54. Treerattanapitak K, Jaruskulchai C (2012) Exponential fuzzy C-means for collaborative filtering. J Comput Sci Technol 27:567–576
55. Verma OP, Katarya R, Bhargava V, Maheshwari N (2011) Use of semantic web in enabling desktop based knowledge management. ICECT 2011–2011 3rd Int Conf Electron Comput Technol 5:190–193
56. Wang F, Qi S, Gao G et al (2016) Logo information recognition in large-scale social media data. Multimed Syst 22:63–73
57. Wu X, Zhu X, Wu G-Q, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26:97–107
58. Xie X, Wang B (2016) Web page recommendation via twofold clustering: considering user behavior and topic relation. Neural Comput Appl 1–9. doi:10.1007/s00521-016-2444-z
59. Yang Y, Yang Y, Shen HT et al (2013) Discriminative nonnegative spectral clustering with out-of-sample extension. IEEE Trans Knowl Data Eng 25:1760–1771
60. Yang Y, Ma Z, Yang Y et al (2015) Multitask spectral clustering by exploring intertask correlation. IEEE Trans Cybern 45:1069–1080
61. Yera R, Castro J, Martínez L (2016) A fuzzy model for managing natural noise in recommender systems. Appl Soft Comput 40:187–198
62. Yu C, Huang L (2016) CluCF: a clustering CF algorithm to address data sparsity problem. Serv Oriented Comput Appl 1–13. doi:10.1007/s11761-016-0191-8
63. Yu X, Liu Y, Huang X, An A (2012) Mining online reviews for predicting sales performance: a case study in the movie domain. IEEE Trans Knowl Data Eng 24:720–734
64. Zhang H-R, Min F (2016) Three-way recommender systems based on random forests. Knowledge-Based Syst 91:275–286
65. Zhang Z, Fang H, Wang H (2016) A new MI-based visualization aided validation index for mining big longitudinal web trial data. IEEE Access 4:2272–2280. doi:10.1109/ACCESS.2016.2569074
66. Zhao WX, Li S, He Y et al (2016) Connecting social media to E-commerce: cold-start product recommendation using microblogging information. IEEE Trans Knowl Data Eng 28:1147–1159
67. Zhu K, Wu R, Ying L, Srikant R (2014) Collaborative filtering with information-rich and information-sparse entities. Mach Learn:177–203

Rahul Katarya is working as an Assistant Professor in Department of Computer Science & Engineering in Delhi Technological University, Delhi India. His research interests include Big Data & prediction, Web Mining Social Networks Recommender systems Personalization, Knowledge discovery, Computational intelligence, Multimedia, Human behaviour, and business intelligence.

Dr. Om Prakash Verma is HOD and Professor in Department of Computer Science & Engineering in Delhi Technological University, Delhi India. His research interests include signal processing image processing soft computing evolutionary computing nature inspired swarm algorithm.