Efficient Fuzzy C-means Based Reduced Feature Set Association Rule Mining Approach for Predicting the User Behavioral Pattern in Web Usage Mining

J. Serin[1], J. SatheeshKumar[2*], T. Amudha[2]

[1] Research Development Center, Bharathiar University, India
[2] Department of Computer Applications, Bharathiar University, India
serin.j@gmail.com, j.satheesh@buc.edu.in, amudhaswamynathan@buc.edu.in

Abstract

Online business and marketing are becoming popular nowadays due to the wide variety of products available from multiple vendors in online. One of the major challenges of e-business merchants is predicting the buying and selling patterns of online customers. Global level competition is another challenge faced by online merchants due to the lowest prices and offers provided by multiple sellers for the same or similar product. Hence, the development of an efficient web mining framework to analyze and predict buyer's interest based on the browsing history will be a great support to the online sellers by providing exact or relevant product details to the buyers in online. Association rule mining plays an essential role in Web Mining for finding the most frequent and predictive patterns of the user. The major challenge in this approach is the generation of many rules for a huge volume of datasets. Decision making based on association rule mining is critical because knowledge is not directly presentin frequent patterns. This research work focuses on the analysis of standard web mining approaches such as k-means clustering, fuzzy c-means clustering, fuzzy k-medoids clustering and fuzzy clustering with weighted session page matrix approach. In this work, MSNBC dataset from UCI Machine Learning Repository has been taken for analysis. Dimensionality reduction plays an important role in the accurate classification of users with respect to their interests. This research work proposed fuzzy C-Means using Kernel Principal Component Analysis (k-PCA) as a dimensionality reduction method based association rule mining classification, grouping and pattern prediction with 100% "confidence" along with a "lift" value greater than 1. The "support" value also shows higher compare with other existing methods and features are effectively reduced in the proposed architecture.

Keywords: Web usage mining, Association rule mining, Fuzzy c-means, Kernel PCA, Pattern discovery

1 Introduction

The World Wide Web continues to be a great source of information gateway due to the rapid increase in the growth of internet usage. The internet users have been increased day by day in which more than 60% of the world's total population is using the internet. The growing rate of internet users from 2000 to 2020 has been drastically increased according to the internet world statistics. Internet usage is rapidly increasing due to various applications such as online shopping and mobile usage. Web Mining involves the discovery of user behavioral patterns from different web servers. The identified user behavioral patterns help organization or merchant to improve the website structure which enhances the availability of information about products for easy access by the user [1]. In today's world, online marketers trying to predict the content to be displayed to their target customers on their website and helps to build customer relationships. To maintain sustainability among the competitors, finding the user's behavioural pattern is a major challenging task. To overcome the research gap, better clustering technique needed to find the hidden user behavioural pattern.

In web usage mining, association rule is needed to find the most frequent behavioral pattern by the web users by using Association Rule Mining [2]. The concept was first introduced by Agarwal et al. to find the frequently purchased items by the customer through market basket analysis [3]. Apriori algorithm is one of the conventional techniques which will generate association rules through candidate set generation. This algorithm shows good performance whenever the support count is low. It is one of the best techniques used in market basket analysis in various fields such as Web Usage Mining (WUM), DNA pattern recognition, stock market analysis and clinical data sets to identify the frequent patterns.

1.1 Motivation of This Research

From the literature work, it is evident that identify the target customers is essential to retain the customers for e-commerce sites. It is also evident that the better clustering will provide more accurate predication for identifying user's behavioral pattern [4-5]. It is inferred from the research work that the target customer is identified by different clustering techniques. In this research work, the web users with similar behavioral browsing pattern are identified through fuzzy c-

*Corresponding Author: J. SatheeshKumar; E-mail: j.satheesh@buc.edu.in
DOI: 10.53106/160792642022122307005

means clustering. The well clustered groups of users are predicted with the next expected group by generating fuzzy inference rules using apriori algorithm. The pages which belongs to the predicted clustered group needs to be improvised for better web personalization.

1.2 Author's Contribution

The author's contribution towards the research work provide well clustered group of pages by using fuzzy c-means clustering technique. During the preprocessing technique, the weighted session pageview matrix has been constructed before applying the clustering technique. In this process, the least frequent pages are effectively removed and only the most frequent pages based on the weight assigned to each page. The kernel principal component analysis technique has been used for dimensionality reduction which helps to find smaller set of variables is known as feature selection and thus the web pages for further mining process are selected through dimensionality reduction techniques. Fuzzy clustering is applied with the reduced set of variables and the web user's behavioral pattern is predicted through the set of fuzzy web inference rules using apriori algorithm to predict the next group of pages which really helps to improve website personalization.

2 Related Works on Predicting User's Behavioral Pattern

Various research works have been proposed in the literature which are used tofind a behavioral pattern of the user in web based applications. Rahul Katarya et al. proposed a novel sequential based approach to implement Fuzzy Clustering for web recommender system using MSNBC dataset [6]. This algorithm was tested with the users who have visited six pages and it predicts the seventh page category that the user is more likely to visit. Boob et al. proposed a work using the fuzzy clustering method to cluster the website browsers into different groups after finding the Session Identification [7]. It enabled the creation of overlapping clusters thereby handling the ambiguity in the data. Guo et al. proposed an algorithm by combining the K-means and fuzzy matrix [8]. This algorithm derives submatrix from the relational matrix based on user and page threshold. This algorithm effectively defines and identifies the number of clusters, initial value and outliers. Eghbal [9] has introduced a fuzzy rule-based clustering algorithm, which explores the potential clusters in the data patterns automatically to identify some interpretable fuzzy rules. Raut et al. proposed a web fuzzy clustering method to be implemented in web user clustering and web page clustering in web usage mining [10]. The web users can be generally classified based on the correlative degree of the web user as firm relation users, hypo-firm relation users, hypo-infirm relation users and infirm relation users. Anjana Gosain et al. have discussed various fuzzy based clustering approaches which were used to discover the frequent pattern by association rule mining and fuzzy c-means clustering to discover inferences [11]. Mangalampalli et al. has proposed a fuzzy ARM algorithm that performs faster than fuzzy apriori for large datasets [12]. Geetharamani et al. has proposed an algorithm called apriori prefix tree algorithm for predicting the subsequent pages visited by the user based on the associative rules which are measured by the metrics such as "support", "lift" and "confidence" [13]. Keivan Kianmehr et al. have proposed a framework by introducing a new layer in the learning process of machine learning while constructing the optimized fuzzy sets and fuzzy association rules [14]. The users with similar behavioral patterns are clustered and then association rules are employed to find the relation and association between the clustered groups. Zhang et al. [15] used fuzzy clustering algorithm to identify the similar users of target groups to improve customer relationship management for internet banking. Sampath et al. [16] implemented an algorithm based on the user's interest to find the frequent pattern mining based on the weight assigned to the page by using systolic tree structure. It reduces the mining time by partitioning the data using FP-growth algorithm. Ashika et al. [17] proposed algorithm to find the user behavioral pattern through improved FP growth algorithm. Mahendra et al. [18] found the customer clusters of similar users of e-commerce sites using k-means clustering algorithm. It is proved that target customers is identified effectively in the buyer's behavior of e-commerce sites.

Xuejun Zhang et al. [19] proposed an architecture for Self-Organizing Map (SOM) use to detect the user profile for online sales through click stream data. They provided recommender system to produce consistent recommendations. Suresh et al. [20] introduced improved fuzzy clustering technique to cluster the similar user's behavioral pattern. They used information entropy to initialize the cluster centers to avoid noisy data in the dataset and their experiment result proved that the clustering is improved with the initial cluster centers. Mobasher et al. [21] presented two techniques about clustering on user transactions as well as clustering on page views in order to identify the aggregate user profiles for web personalization based on the calculation of recommendation score.

Finally, based on the literature work the user's behavioral pattern is evaluated either through clustering or rule based mining approach. In the proposed research work, both approaches are hybrid to attain the effective targeted clusters group and then predict the user's behavioral pattern through fuzzy web inference rules by implementing apriori algorithm. The proposed architecture helps to improve the effective web portal management.

3 Fuzzy C-means and Association Rule Mining

Various clustering and mining approaches have been proposed in existing literatures. Most of these algorithms struggle to classify the web user's behaviors [22]. Hence, predicting user interest is a great challenge to the web mining researchers. Standard methods used for map ping user's interest with respect to the web page include k-means clustering, fuzzy c-means clustering, fuzzy k-medoids clustering, fuzzy clustering with weighted session page matrix approach and fuzzy with kernel PCA method. The primary advantage of fuzzy c-means clustering is that the algorithm correctly segregates the user interest which might be in more

than one page. The frequent if-then association pattern has been identified by using association rules in the clustered user groups.

3.1 K-Means Clustering

K-means clustering is an unsupervised learning algorithm to group the unlabeled dataset into different groups and each data point belongs to only one group which has similar properties. It is one of the easiest techniques for creating groups by optimizing the criterion function [14]. This algorithm was implemented into two phases. In the first phase, k-centroids for each cluster have been determined and in the second phase, each data point is associated with the nearest centroid. New centroids were calculated till the centroids won't move anymore. A major drawback of this algorithm is selection of initial centroids and it is also computationally expensive which requires the time proportional to the number of data items and number of clusters [23].

3.2 Fuzzy C-means (FCM) Clustering

In traditional clustering algorithms such as K-means and K-mediods are partitioning the data objects into distinct clusters where each data points belong to only one cluster. Instead, in Fuzzy C-means clustering algorithm, the data points belong to more than one cluster and the concept was first introduced by Dunn [24] and then improved by Bezdek [25]. The user who is interested in sports may be interested in health and technology as well. Hence, a single user may fit into more than one group. This minimizes dissimilarities between objects for the fuzzy clustering algorithm. The objective function used for minimizing the dissimilarity by FCM is given by Equation 1.

$$J_m = \sum_{i=1}^{c} \sum_{j=1}^{m} q_{ij}^{n} * d_{ij}^{2}, \quad 1 \leq n < \infty \qquad (1)$$

where $d_{ij} = x_j - c_j^2$, 'Q_{ij}' is the degree of membership function matrix, 'C' is the number of clusters, 'm' represents the number of data records, 'N' is any real number greater than 1, 'd_{ij}' is the distance from 'x_j' to 'c_j' ie. 'c_j' denotes the cluster center of the j^{th} cluster to the i^{th} iteration.

3.3 Fuzzy K-medoids Clustering

Fuzzy k-medoids is a soft clustering technique in which the data points showa high degree of similarity when partitioning a set of objects into 'k' clusters. This method searches for k-medoids which minimize the dissimilarity of all the objects in the dataset to the nearest medoid. These medoids to be calculated when mean or centroids cannot be defined in the dataset such as gene expression or 3-D trajectories [26]. The K-medoids algorithm has been implemented by two phases such as object selection and clustering. In the first phase, initial clustering is obtained by successive selection of objects until 'k' objects have been found. In the second phase, a set of representative objects has attempted to improve clustering. The function for dissimilarity of each non-medoid (f_i) with the medoid object (d_i) is calculated by using the Manhattan distance measure as in Equation 2.

$$c = \sum_{fi} \sum_{di \epsilon fi} di - fi. \qquad (2)$$

The major drawback of this algorithm is that it is less sensitive to outliers than other clustering algorithms and it is not suitable for non-spherical group of objects.

3.4 Fuzzy Clustering with Weighted Session Page View Matrix

Log files obtained from the web server needs different preprocessing cycles since the raw information in the log file are not suitable for processing. Each row represents the user and the column represents the frequency of the pages visited by the user otherwise it is mentioned as zero. Weight of a page has been calculated by the number of times that page is visited with respect to a total number of pages in each session and the pages with greater threshold value have been retained and the remaining pages will be removed. The weight can be assigned by using the formula given in equation 3.

$$W_{ij} = \frac{m}{\sum_{i=1}^{n} np}, \qquad (3)$$

where m is the frequency of each page and np is the total number of page requests in a session and then construct the user session matrix to perform normalization of the data.

3.5 Fuzzy with Kernel Principal Component Analysis

A major challenge triggered in the field of web usage mining is the number of pages in the website which is to be considered as attributes. To overcome these challenges, dimensionality reduction is one of the major steps in preprocessing phase which plays a vital role in analyzing the log files. It is the process of reducing number of pages with respect to its closeness to improve the quality of data. It reduces the attribute features either by combining, merging or ignoring irrelevant features from the input feature set in such a way that it will not lose the significant characteristics in the original dataset [27]. In kernel-Principal Component Analysis (KPCA), the nonlinear function is much needed to reduce the dimensionality of a dataset for linearly inseparable data into linearly separable data by projecting them into high dimensional space. Mapping function is required to project the data from a lower dimensional space to a higher dimensional space to make the dataset into linearly separable [28].

4 Efficient Fuzzy C-means Clustering Based ARM Architecture for Web User Pattern Analysis

This research work presents a novel method to inference knowledge system for association rule mining. Initially, the user log files are taken from the web server. After preprocessing phase, the data are transformed into numerals for further processing. Next, the users are clustered into groups by applying the Fuzzy C-means algorithm. The resultant data

in the form of a fuzzy membership matrix obtained by using fuzzy C-means algorithm are converted into transaction data. Cluster data will be classified based on fuzzy web inference. The fuzzy membership matrix in the form of transaction data has been applied to Apriori algorithm to generate association rules. The generated association rules for the clustered group predict the behavioral pattern of the user which will help to improve the web personalization. Now days, even single start-up companies delivering sophisticated website content to their users and their focus is to create instinctive web experiences for each web users. For example, Netflix has a knack for delivering the favourite movies based on their previous user's visits of the website. Therefore, our proposed research work forms the clustering of similar users and predict the user behavioural pattern which really helps to enhance their topical content marketing, advertising and retail. The activity diagram of the proposed work is given in Figure 1.

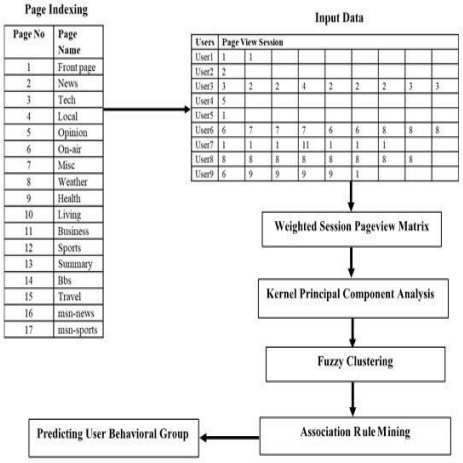

Figure 1. Architecture of fuzzy clustering with association rule mining

Normally, users visit more than one page in the website. Hence, fuzzy clustering is used to group similar user's browsing behavior. In this example, the optimal number of clusters is found by elbow method and it is identified to be five. With the help of clustering output, the fuzzy web inference rules have been generated. This step matches and identifies the similar user group based on their browsing pattern. Then, the association rule mining has been used to predict the user group [29].

4.1 Dimensionality Reduction Based Fuzzy Inference System using ARM

Even though, the efficient framework proposed in Figure 1 performs better than existing methods, number of relevant and irrelevant attributes decides the prediction accuracy of the proposed system. Hence, this phase of research work includes one of the efficient dimensionality reduction methods as part of the proposed architecture. Figure 2 shows the enhanced version of fuzzy clustering through weighted session page view matrix as well as kernel PCA based dimensionality reduction technique used in proposed approach. The weighted session page view matrix is constructed to improve the accuracy of the clustering results improves clustering accuracy [30]. Dimensionality reduction removes redundant features, noise and also extracted the featured data which leads to less computation time and occupies less storage. The K-PCA technique has been used to identify more contributing pages in the website which helped to improve the quality of data and thereby.

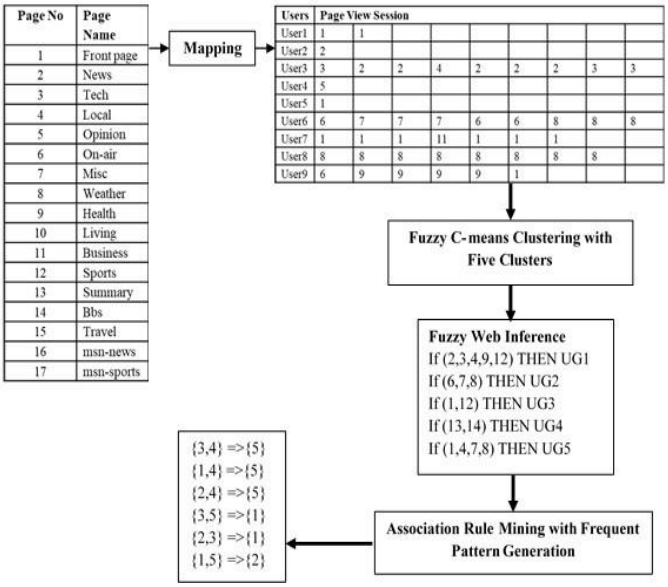

Figure 2. Enhanced fuzzy clustering with association rule mining through weighted session page view matrix and K-PCA reduction method

5 Materials and Methods

5.1 Data Set

The weblog files of the "msnbc.com" web site from UCI KDD archive at the University of California have been used for an experimental purpose [31]. The log files were generated from Internet Information Server (IIS). Each page in the website is represented as an integer value for the smooth processing of data. It contains the news-related portions of "msnbc.com" for the entire day of 28th September 1999. Each row in the dataset corresponds to a page viewof a user during their browsing period. Each page is represented as an integer and each row represents the hits of a single user. The sample dataset is given in Table 1.

Table 1. Sample MSN dataset

User no.	Page view session									
User 1	3	2	2	4	2	2	2	3	3	
User 2	6	7	7	7	6	6	8	8	8	8
User 3	6	9	4	4	4	10	3	10	5	10
User 4	1	1	10	1	1	1				
User 5	8	8	8	8	8					

5.2 Experimental Analysis

The raw data is preprocessed to improve the quality of data to apply mining techniques and hence the clustering performance has been improved in parallel. Table 2 shows that, User 1 visited the "News" page five times, "Tech" page thrice and "local" page one time. In Table 2, each row represents the pages visited by the user and each column represents the pages

presented in the website as attributes. Column 1 represents "front page", Column 2 represents "News", Column 3 represents "Tech" and so on and thereby first user hits "News" page five times, "tech" page three times and "local" page one time. The second user hits "On-air" thrice, "Misc" page thrice and "Weather" page four times. The weighted session page view matrix has been constructed as shown in Table 3. The step-by-step implementation of proposed framework is shown in Figure 3. The weights of each column have been calculated by dividing frequency of the page with a total number of pages in each session.

Table 2. Preprocessed MSN log file

Users	Front page	News	Tech	Local	Opinion	On-air	Misc	Weather	Health	Living	...	msn-sports
User 1	0	5	3	1	0	0	0	0	0	0	..	0
User 2	0	0	0	0	0	3	3	4	0	0	..	0
User 3	0	0	1	3	1	1	0	0	1	3	..	0
User 4	5	0	0	0	0	0	0	0	0	1	..	0
User 5	0	0	0	0	0	0	0	5	0	0	..	0

Table 3. Weighted session page view matrix

Front page	News	Tech	Local	Opinion	On-air	Misc	Weather	Health	Living	Business	Sports	Summary	BBs	Travel	msn-news	msn-sports
0	0.55	0.33	1.99	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0.3	0.3	0.4	0	0	0	0	0	0	0	0	0
0	0	0.1	0.3	0.1	0.1	0	0	0.1	0.3	0	0	0	0	0	0	0
0.83	0	0	0	0	0	0	0	0	0	0.16	0	0	0	0	0	0

The K-PCA dimensionality reduction technique has been used to identify the contributing attributes towards further mining techniques. Fuzzy clustering has been applied to the MSN data set to cluster the similar browser behavioral patterns. The optimal number of clusters is identified by using NbClust method in R for the given dataset. The output of clustered membership matrix grouped into five clusters by using the optimal number of clusters as shown in Table 4.

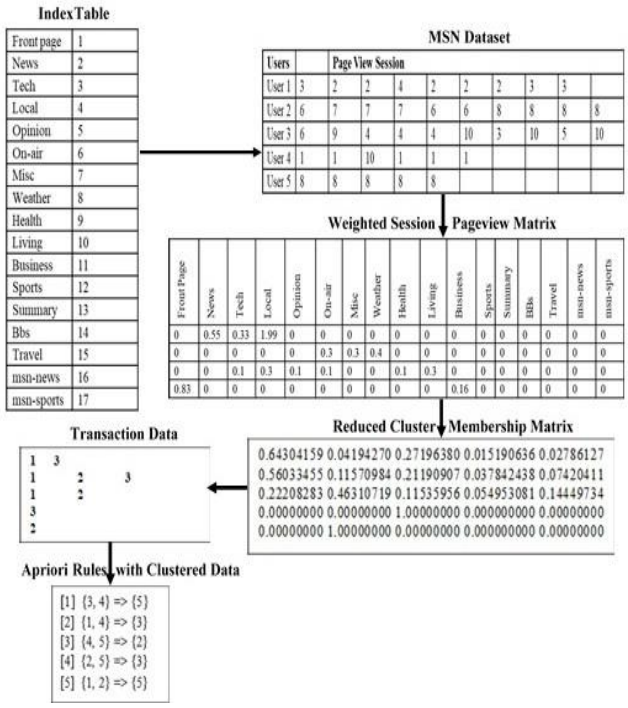

Figure 3. Results of proposed fuzzy c-means based reduced feature set clustering architecture

In this membership matrix, each row represents every user visit and each column value in that row represents the probability of being in that cluster.The user group can be identified by taking a sample of 100 records and find the clustered group. The Fuzzy web inference rules have been generated by analyzing the cluster as shown in Figure 4. Fuzzy inference rule can be mapped to the given dataset as represented in Table 5. Five different user groups have been created by the proposed method after applying fuzzy cluster. The fuzzy inference rule is constructed by segregating the user groups based on the similarity of behavioral patterns. The Group1 users are interested in "News", "tech", "local" and "health" pages and the second group is interested in the pages such as "On-air", "Misc" and "weather" pages. Each row in the membership matrix is converted into transaction data to apply an apriori algorithm in order to predict the user's behavioral pattern. For example, the membership value and the corresponding transaction datafor the first row are shown in Table 6. The membership value greater than 0.1 is retained with the corresponding clustering number and the remaining clusters are not taken into consideration for the apriori algorithm. Association rule mining has been applied to the clustered group in order to avoid confusion raised in generating the rules for making a decision. The sample association rules with the lengths two with "confidence" level above 85% have shown in Table 7.

Pages in (2, 3, 4, 9)	UG1
Pages in (6, 7, 8)	UG2
Pages in (1, 12)	UG3
Pages in (13, 14)	UG4
Pages in (1, 4, 7, 8)	UG5

Figure 4. Fuzzy inference rules

Table 4. Membership matrix of five clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
0.64304159	0.04194270	0.27196380	0.015190636	0.02786127
0.56033455	0.11570984	0.21190907	0.037842438	0.07420411
0.22208283	0.46310719	0.11535956	0.054953081	0.14449734
0.00000000	0.00000000	1.00000000	0.000000000	0.00000000
0.00000000	1.00000000	0.00000000	0.000000000	0.00000000

Table 5. Fuzzy inference results

User Group 1	News	Tech	Local	Health
User Group 2	On-air	Misc	Weather	-
User Group 3	Front page	Sports	-	-
User Group 4	Summary	BBs	-	-
User Group 5	Weather	Front page	Misc	-

Table 6. Membership value vs transaction data

Membership value	0.64304159	0.04194270	0.2719638	0.01519063	0.0278612
Transaction data	1	-	3	-	

Table 7. Association rules

Rules	Support	Confidence	Lift
{3, 4} =>{5}	0.5951265	1.0000000	2.1045365
{1, 4} =>{3}	0.8716026	1.0000000	2.104536
{4, 5} =>{2}	0.62135895	1.0000000	2.104536
{2, 5} =>{4}	0.732951265	1.0000000	1.948858
{4, 2} =>{1}	0.79151265	1.0000000	1.912186

5.3 Performance Analysis Metrics

Silhouette is one of the measures used to calculate the goodness metric of a cluster and this metric also estimates the average distance between the clusters.

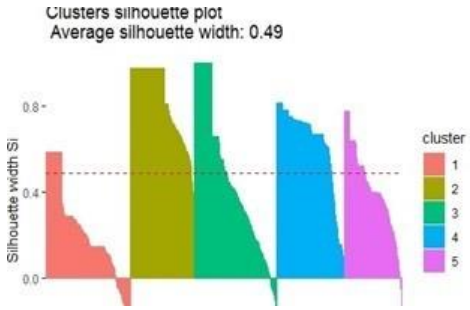

Figure 5. Silhouette width for five clusters

The silhouette width is calculated by using the Equation 4, where, A_i is an average dissimilarity between 'i' and all other points and C_i is the minimum value of dissimilarity between 'i' and its neighbor cluster. The silhouette plot shown in Figure 5 represents a measure of closeness between each point and its neighboring clusters. In this plot diagram, x-axis denotes the number of clusters and y-axis denotes the silhouette width. The silhouette widths of each cluster is shown in Table 8.

$$\text{Si} = \frac{(C_i - A_i)}{\max(A_i, C_i)}. \quad \textbf{(4)}$$

Table 8. Silhouette width

Cluster no	Width
1	0.56
2	0.94
3	0.96
4	0.78
5	0.73

6 Results and Discussion

The results of the proposed system have been analyzed to find the possibilities of a recommended user group to improve the web portal management using 2,500 records of MSN website. The silhouette analysis for each cluster interprets measurement of observation between the clusters and it also estimates average distance between the clusters. The silhouette widths between the clusters have been measured using k-means, fuzzy c-means and fuzzy k-medoids to determine the better clustering environment. Table 9 shows the resultant value of silhouette width obtained using proposed weighted session page view matrix with the K-PCA dimensionality reduction technique. The table clearly shows that enhanced fuzzy c-means with weighted session page view matrix and K-PCA techniques perform the better clustering.

Table 9. Average silhouette width based on weighted session and kernel PCA

Algorithm	Without weighted session Page view & K-PCA	With weighted session Page view & K-PCA
K-means	0.52	0.67
Fuzzy C-means	0.74	0.87
Fuzzy K-medoids	0.63	0.79

Dimensionality reduction plays an important role in accurate classification of user's behavior patterns. In this research work, one of the better dimensionality reduction methods such as Kernel PCA has been applied to extract the features in the given dataset in order to improve the quality of clustering. Fuzzy C-means clustering has been applied on the reduced feature set to identify similar user groups. Association rule mining has been used to predict the user's behavioral pattern for web personalization. The association rules generated by the apriori algorithm produces enormous rules and it is difficult for the researchers to find interesting patterns. Initially, ARM applied on this dataset generates 22,741 rules which are highly difficult to take a decision. Fuzzy C-means without dimensionality reduction-based ARM generates 102 rules which are better than association rule mining approach [32]. But still, the clustering accuracy needs to be increased for better prediction of user behaviors. Hence, to improve the quality of clustering, kernel PCA based dimensionality reduction has been applied in proposed architecture based on Fuzzy C-means clustering along with association rule mining. Results show that the proposed architecture generates reduced number of rules as 89 which is better than other the two approaches. The comparative results of all three approaches have shown in Figure 6.

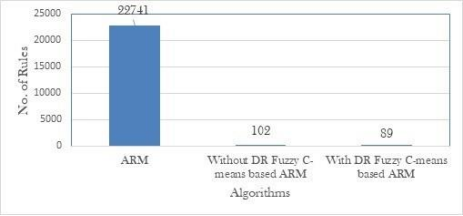

Figure 6. Number of rules generated by ARM, fuzzy c-means based ARM and fuzzy c-means based ARM with dimensionality reduction

6.1 Analysis of Fuzzy Association Rules

The association rules generated by the apriori algorithm finds some interesting relations or associations in between them. Each rule is analysed by the proposed framework states that the more frequently group of pages visited by the user in 3 and 4 is more likely to visit the group five pages with the support and confidence metric. Hence, if a user visits group 3 and 4 also visits group 5 with cent percent confidence. Therefore, the administrator of the website owner focus to enhance the pages in group 5 for better serve the needs of the user.

The statistical measures such as "support", "confidence" and "lift" are benchmarking methods used to assess the performance of proposed architecture. The "support" value indicates that how frequently user visits the pages. The "confidence" value measures number of times the relationship exists between pages to be true. The "lift" metric helps to determine how many times the fuzzy association rules are expected to be true and it also measures the correlation that exists between the pages. If the "lift" value is equal to 1 then the probability between pages is independent of each other. If the value is greater than 1 then the relationship between the item set is dependent on each other. Hence, the ratio of the higher "lift" value means that the itemsets appear together more than the expected. Table 10 represents the sample fuzzy k-medoids association rules generated after clustering the similar browsing behavioral users along with the "confidence" and "lift" measurements.

Table 10. Fuzzy K-medoids association rules

Rules	Support	Confidence	Lift
{3, 4} =>5	0.0595	0.39318	0.8324
{1, 3} =>5	0.1029	0.32308	0.7489
{2, 5} =>3	0.1078	0.2973	0.68919
{4, 5} =>2	0.1176	0.39344	0.72699
{2, 5} =>1	0.1862	0.43182	0.91761

The association rules generated by this algorithm executed with less "lift" value. The "lift" value, which is less than 1, indicates that the clustering group is independent of each other. Table 11 indicates that the association rules generated by fuzzy c-means algorithm with little higher "lift" values. Table 12 denotes the fuzzy association rules generated by the proposed framework and the results proved 100% of confidence values along with high "lift" value. The results indicate that cluster groups 3 and 4 are supposed to move to 5^{th} cluster group. Hence, the pages in cluster group 5 need to be the modified for web personalization. The average values taken for the performance metrics of association rules are given in line chart as shown in Figure 7 and the experimental result shows that the performance metrics of the association rules is high in the proposed research work. Figure 8 shows the predicted cluster groups of each user group. The X-axis represents predicted user group and Y-axis represents the cluster user group. User groups belong to 4 and 2 have predicted to be in cluster 1 and user groups 4 & 5 predicted to be in cluster 2 and the users in cluster groups 1 and 4 have predicted to be in cluster 3 and so on. For example, users who are reading "front page", "sports" and "bbs" will be slightly interested to read "weather" page. Hence, "weather" page needs to be modified and the website to be more personalized based on the user's need by using the proposed framework.

Table 11. Fuzzy C-means association rules

Rules	Support	Confidence	Lift
{3, 4} =>5	0.087	0.4386	0.8549
{2, 4} =>3	0.0595	0.3931	0.8324
{4, 5} =>1	0.0871	0.4386	0.8549
{1, 3} =>5	0.1991	0.4761	0.9927
{1, 3} =>2	0.2412	0.5723	1.095

Table 12. Enhanced fuzzy association rules using the proposed architecture

Rules	Support	Confidence	Lift
{3, 4} =>{5}	0.59512	1.00000	2.10453
{1, 4} =>{3}	0.87160	1.00000	2.10456
{4, 5} =>{2}	0.62135	1.00000	2.10453
{3, 5} =>{1}	0.73295	1.00000	1.94885
{1, 3} =>{2}	0.72345	1.00000	1.92100
{4, 2} =>{1}	0.79151	1.00000	1.91218

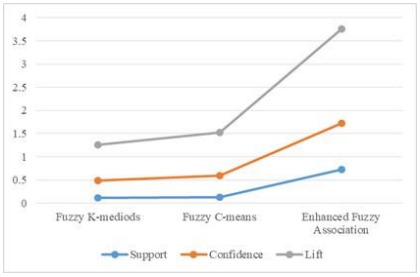

Figure 7. Average values of performance metrics

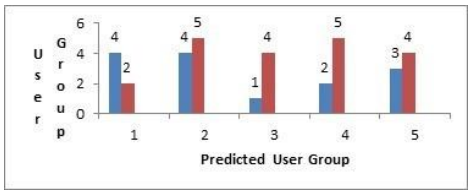

Figure 8. User vs predicted cluster group

7 Conclusion

In this research work, a novel framework using Fuzzy C-means based Inference system for association rule mining has been proposed to find the user behavioral pattern in web usage mining. The proposed method also includes kernel-PCA as a dimensionality reduction method for the best clustering of users with respect to their interest. The proposed architecture has three phases such as preprocessing, clustering and predicted mapping. In the first phase, the data obtained from web server is preprocessed and converted into numerals for smooth processing. In the second phase, Fuzzy C-means clustering has been applied using a reduced feature set to cluster the similar user interest behavioral characteristics and find the fuzzy inference by doing the cluster analysis. In the final phase, accurate mapping of users to the respective groups have been done. Results show that proposed algorithm works better than existing standard approaches. However, this proposed research work provides new insights and challenges need to be solved is to find the better clustering techniques. The apriori algorithm helps to extract interesting rules increases the number of database scan and also the number of candidates set generation also increases. Finding the better

technique of generation of association rule algorithm which removes the disadvantages of apriori algorithm.

Acknowledgement

Authors J. Serin, J. SatheeshKumar and T. Amudha are equally contributed in this research work.

References

[1] G. Navarro-Arribas, V. Torra, Towards microaggregation of log files for Web usage mining in B2C e-commerce, *International Conference on North American Fuzzy Information Processing Society*, Cincinnati, OH, USA, 2009, pp. 1-6.

[2] F. M. Facca, P. L. Lanzi, Mining interesting knowledge from weblogs: a survey, *Data & Knowledge Engineering*, Vol. 53, No. 3, pp. 225-241, June, 2005.

[3] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, *International conference on Management of data*, Washington, DC, USA, 1993, pp. 207-216.

[4] O. Zaiane, *Web usage mining for a better web-based learning environment*, pp. 1-5, 2001. https://webdocs.cs.ualberta.ca/~zaiane/postscript/CATE2001.pdf

[5] H. Zhang, W. Liang, An intelligent algorithm of data pre-processing in Web usage mining, *Fifth World congress on Intelligent Control and Automation*, Hangzhou, China, 2004, pp. 3119-3123.

[6] R. Katarya, O. P. Verma, An effective web page recommender system with fuzzy c-mean clustering, *Multimedia Tools and Applications*, Vol. 76, No. 20, pp. 21481-21496, October, 2017.

[7] A. N. Boob, D. M. Dakhane, Fuzzy Clustering: An Approach for Mining Usage Profiles from Web, *International Journal of Computer Technology and Applications*, Vol. 3, No. 1, pp. 329-331, 2012.

[8] J. Guo, S. Zhang, Z. Qiu, *Efficient K-means clustering algorithm in web log mining*, 2016. https://www.researchgate.net/publication/309529379_Efficient_K-Means_Clustering_Algorithm_in_Web_Log_Mining

[9] E. G. Mansoori, FRBC: A fuzzy rule-based clustering algorithm, *Institute of Electrical and Electronics Engineers transactions on fuzzy systems*, Vol. 19, No. 5, pp. 960-971, October, 2011.

[10] A. B. Raut, G. R. Bamnote, Web document clustering using fuzzy equivalence relations, *Journal of Emerging Trends in Computing and Information Sciences, CIS Journal*, Vol. 2, pp. 22-27, 2010.

[11] A. Gosain, S. Dahiya, Performance analysis of various fuzzy clustering algorithms: a review, *Procedia Computer Science*, Vol. 79, pp. 100-111, 2016.

[12] A. Mangalampalli, V. Pudi, FAR-miner: a fast and efficient algorithm for fuzzy association rule mining, *International Journal of Business Intelligence and Data Mining*, Vol. 7, No. 4, pp. 288-317, January, 2012.

[13] R. Geetharamani, P. Revathy, S. G. Jacob, Prediction of users webpage access behaviour using association rule mining, *Sadhana*, Vol. 40, No. 8, pp. 2353-2365, December, 2015.

[14] J. Wu, The uniform effect of k-means clustering, in: Advances in K-means Clustering, *Springer*, Berlin, Heidelberg, 2012, pp. 17-35.

[15] J. Zhang, P. Zhao, L. Shang, L. Wang, Web usage mining based on fuzzy clustering in identifying target group, *International Colloquium on Computing, Communication, Control, and Management*, Sanya, China, 2009, pp. 209-212.

[16] P. Sampath, C. Ramesh, T. Kalaiyarasi, S. S. Banu, G. A. Selvan, An efficient weighted rule mining for web logs using systolic tree, *IEEE-International Conference on Advances In Engineering, Science And Management*, *Nagapattinam, India*, 2012, pp. 432-436.

[17] A. Gupta, R. Arora, R. Sikarwar, N. Saxena, Web usage mining using improved Frequent Pattern Tree algorithms, *International Conference on Issues and Challenges in Intelligent Computing Techniques*, Ghaziabad, India, 2014, pp. 573-578.

[18] M. P. Yadav, M. Feeroz, V. K. Yadav, Mining the customer behavior using web usage mining in e-commerce, *2014 IEEE International Conference on Computing, Communication and Networking Technologies*, Karur, Tamil Nādu, 2012, pp. 1-5.

[19] X. Zhang, J. Edwards, J. Harding, Personalised online sales using web usage data mining, *Computers in Industry*, Vol. 58, No. 8-9, pp. 772-782, December, 2007.

[20] K. Suresh, R. M. Mohana, A. R. M. Reddy, A. Subramanyam, Improved FCM algorithm for clustering on web usage mining, *IEEE International Conference on Computer and Management (CAMAN)*, Wuhan, China, 2011, pp. 1-4.

[21] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Discovery and evaluation of aggregate usage profiles for web personalization, *Data mining and knowledge discovery*, Vol. 6, No. 1, pp. 61-82, January, 2002.

[22] K. Santhisree, A. Damodaram, CLIQUE: Clustering based on density on web usage data: Experiments and test results, *International Conference on Electronics Computer Technology*, Kanyakumari, India, 2011, pp. 233-236.

[23] C. Bouras, V. Tsogkas, Improving news articles recommendations via user clustering, *International Journal of Machine Learning and Cybernetics*, Vol. 8, No. 1, pp. 223-237, February, 2017.

[24] J. C. Dunn, Well-separated clusters and optimal fuzzy partitions, *Journal of cybernetics*, Vol. 4, No. 1, pp. 95-104, 1974.

[25] J. C. Bezdek, Pattern recognition with fuzzy objective function algorithms, *Springer Science & Business Media*, 2013, pp. 1-255.

[26] L. Kaufman, P. Rousseeuw, Clustering by means of medoids, in: Y. Dodge (Eds.), *statistical data analysis based on the l1-norm and related methods*, 1987, pp. 405-416.

[27] K. I. Kim, K. Jung, H. J. Kim, Face recognition using kernel principal component analysis, *Institute of Electrical and Electronics Engineers signal processing letters*, Vol. 9, No. 2, pp. 40-42, February, 2002.

[28] Z. Liu, D. Chen, H. Bensmail, Gene expression data classification with kernel principal component

analysis, *Journal of Biomedicine and Biotechnology*, Vol. 2005, No. 2, pp. 155-159, 2005.

[29] G. Pallis, L. Angelis, A. Vakali, Validation and interpretation of Web users' sessions clusters, *Information processing & management*, Vol. 43, No. 5, pp. 1348-1367, September, 2007.

[30] T. W. Cheng, D. B. Goldgof, L. Hall, Fast clustering with application to fuzzy rule generation, *4th IEEE International Conference on Fuzzy Systems*, Yokohama, Japan, 1995, pp. 2289-2295.

[31] A. Pitman, M. Zanker, M. Fuchs, M. Lexhagen, in: U. Gretzel, R. Law, M. Fuchs (Eds.), Web Usage Mining in Tourism — A Query Term Analysis and Clustering Approach, *Information and Communication Technologies in Tourism*, Springer, 2010, pp. 393-403.

[32] S. Araya, M. Silva, R. Weber, A methodology for web usage mining and its application to target group identification, *Fuzzy sets and systems*, Vol. 148, No. 1, 139-152, November, 2004.

Biographies

J. Serin, received a B. Sc Degree in Computer Science in 2001, Master of Computer Applications Degree in 2004 and Master of Philosophy in Computer Science Degree from Madurai Kamaraj University, Madurai, in 2007. She is currently a Ph.D scholar in Bharathiyar University, Coimbatore. Her research interests include Web Mining and Machine Learning.

J. SatheeshKumar is serving as Associate Professor in the Department of Computer Applications, Bharathiar University. He received an Appreciation Award by Korean Convergence Society, Korea. He has 21 years of teaching and 15 years of research experience. He is a professional member of CSI, IEEE, IEEE System Man and Cybernetics, ACM, IET, GVIP, ACEEE and Hikey Media. He can be contacted through j.satheesh@buc.edu.in, jsathee@rediffmail.com.

T. Amudha is serving as Associate Professor in the Department of Computer Applications, Bharathiar University. She has over 21 years of teaching and 15 years of research experience. She is actively engaged in Teaching and Research in Artificial Intelligence, Bio-inspired Optimization algorithms, Swarm Intelligence and Software Agents. She is a life member of Computer Society of India and professional member of ACM, IEEE, ISCA and IAENG. She can be connected through amudhaswamynathan@buc.edu.in, amudha.swamynathan@gmail.com.