Indian Journal of Science and Technology, *Vol 8(12), 69283, June 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Clustering of Navigation Patterns using Bolzwano_Weierstrass Theorem

V. Chitraa[1*] and Antony Selvadoss Thanamani[2]

[1]CMS College of Science & Commerce (Autonomous), Coimbatore, Tamilnadu, India; vchit@yahoo.co.in
[2]Computer Science, NGM College (Autonomous), Pollachi, Coimbatore, Tamilnadu, India; selvdoss@gmail.com

Abstract

Objectives: The primary objective of this research paper is to design a new and efficient clustering technique to group user navigation patterns which are useful for classification system to classify a new user with the previous users group. **Methodology:** Three real time web log data sets are collected from e-commerce web server, academic institution web server and a research journal web server. All three sets were collected from IIS web servers. After navigation patterns are derived from preprocessing step it is clustered into groups by using traditional Fuzzy C-Means technique. The clusters are validated and re-clustered using Bolzano_Weierstrass Theorem. **Findings:** Web log data is preprocessed and ICA is applied in the user session matrix to select relevant and important features. To measure the clustering accuracy of proposed and the existing methods, the parameters such as Rand Index, F measure are calculated and compared. It shows proposed BWFCM have higher rand index rate than FCM and lesser error rate. To understand the impact of the feature selection method, the data sets were implemented with the existing and proposed methods of feature selection. The parameters taken for comparison were Rand Index, Sum of Squared Errors, F-measure. The method was implemented in all the three data sets after data cleaning, session construction step. Clustering was carried out twice with the proposed clustering algorithm in all the three data sets, without selecting features and after selecting features. It was observed that the clustering results are poor when applied in full data set with irrelevant features, and the performance was increased after relevant features were selected. **Conclusion:** The result of the optimized clustering proves its significance and there is an increase in similarity of intra clustering and dissimilarity in inter clustering than the existing methods.

Keywords: Bolzano_Weierstrass Theorem, Clustering, Feature Selection, Navigation Patterns, Web Usage Mining

1. Introduction

Web Usage Mining is an emerging technique where mining techniques are applied in web log data which are collected from web server. It is categorized as implicit mining since analysis is done without the knowledge of users. User behavior analysis is one of the purposeful analysis in web usage mining which is used in different domains such as e-commerce, e-learning, health sectors to improve the system and to evolve the system design as per user's interest, etc. It is also used in Business analytics which help an organization to prepare for future growth and challenges.

1.1 Phases of Web Usage Mining

The mining process is carried out in three phases such as preprocessing, pattern discovery and pattern analysis.

Preprocessing is an important task to improve the quality and it impacts the resultant rules and models produced by the data mining algorithm. The objective of preprocessing is to transform the raw log data into a meaningful set of user profiles[1]. It selects standardized data from the original log files, prepared for user navigation pattern discovery algorithm and it is time consuming by taking almost 80% of mining process[2]. The preprocessed log data directly affects the accuracy and reliability of the algorithm processing results[3].

**Address for correspondence:*

The second important phase is pattern discovery in which different data mining techniques like statistical analysis, association, clustering, pattern matching etc., are used to process the data. Clustering is the process of grouping data into many groups of similar objects. Informally, clustering is also considered as data modeling since it concisely summarize the data and relates to many disciplines from statistics to numerical analysis. User's navigation patterns are clustered into groups to find user groups who have common interests based on their behaviors. Clustering is different in log data due to its nature from traditional clustering. Another technique is classification where data is mapped into one of several predefined classes. A learning model is build which is used to classify a class of objects to determine the class label of a new object whose class is not determined. The process is termed as supervised learning because the class label of each training sample is provided. The classification of web usage data is usually used to construct profiles of users belonging to a particular class or category.

Pattern Analysis is the final phase where the discovered patterns are further processed and filtered producing models that can be served as an input to different visualization tools and report generation tools. This paper mainly focuses on the grouping of selected features by Bolzano_Weierstrass based clustering of patterns.

This research paper is segregated into five sections and structured as follows. Section two analyzes works related to this paper. Proposed Methodology is described in section three. Section four presents experimental results and the last section concludes the whole process.

2. Related Work

For effective clustering reduction in dimension is essential to make the high dimensional data addressable and reduces the computational cost, and also provide users with a clear picture and visual examination of the interesting data. Many emerging dimensionality reduction techniques have been proposed in the literature. For example, Local Dimensionality Reduction (LDR) approach tries to find local correlations in the data, and performs dimensionality reduction on the locally correlated clusters of individual data[4] and dimensionality reduction adaptively adjusted and integrated with the clustering process[5].

To identify the optimal user profile from the given web usage data, a Simulated Annealing is used as an optimization tool for biclustering of web usage data. Extracted biclusters consists of correlated users whose usage behaviors are similar across the subset of web pages of a web site where as these users are uncorrelated for remaining pages of a web site[6]. The simplest and most used by researchers is Fuzzy c-means method otherwise known as soft clustering which allows one piece of data to belong to two or more clusters. They observed the clustering is found to be very much useful for clustering web log data due to the fact that users' interest may vary from time to time and the browsing pattern will vary. Two factors such as page-click number and web browsing time stored in the web log are considered due to its fuzzy and uncertain behavior[7]. The author describes the concept of time discretization and applies fuzzy equivalence relation clustering to classify web users[8]. An improved version of FCM is proposed for large and noisy data set[9]. User access patterns with similar surfing behaviors are clustered into one class. Time duration shows interest and denoted by fuzzy variable and access patterns are characterized by fuzzy vector. Similarity is applied and a rough approximation based clustering approach is adopted to cluster. Rough k-means clustering on fuzzy web access patterns is proposed to cluster web access patterns[10]. To overcome the noise problem in FCM a hybrid model of Possibilistic C means and Fuzzy C Means is proposed in which Possibilistic type of membership function is considered by modifying the constraint condition of FCM[11]. The advantages of both fuzzy and possibilistic c-means techniques is merged. Memberships and typicality's are essential for the accurate characteristic of data substructure in clustering technique. The authors modified[12] FPCM to form a new objective function with a weighting exponent and grouped the user sessions effectively and proved the increase in accuracy and decrease in time and error rate. Rough-fuzzy c-means, is proposed which is a hybrid unsupervised learning algorithm and comprises a judicious integration of the principles of rough sets and fuzzy sets. The membership function of fuzzy sets enables efficient handling of overlapping partitions while the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in class definition. A constraint to improve FPCM is introduced by Vanisri and Loganathan[13]. Possibilistic reasoning strategy on fuzzy clustering with penalized and compensated constraints for updating the grades of membership and typicality to overcome the problems in FPCM. The authors define user sessions and discuss clustering ses-

sions based on the pair-wise dissimilarities using a robust fuzzy clustering algorithm in[14]. Fuzzy Artificial Immune System and clustering techniques are coupled to improve the users' profiles obtained through clustering[15]. FCM is combined with ant based clustering and the model is employed for clustering Web users based on their accesses to Web pages[16]. A new algorithm called t-bridge is applied to fuzzy equivalence matrix clustering algorithm[17] and it is an alternative for finding transitive closure of matrix which is difficult. The sessions are formed as a matrix and transformed into web fuzzy matrix and clustering is done based on relative active degree of the web users and correlative degree of web users. Cluster validity index is used for rough fuzzy c-means clustering algorithm which roots on probabilistic metric called rough fuzzy Bayesian like validation method. Maximum Bayesian score stipulates optimal number of cluster[18]. Navigation patterns are discovered by density based clustering algorithm and an online navigation pattern prediction is proposed by use of K nearest neighbor algorithm along with inverted index concept. The prediction accuracy of patterns is increased by modifying the pattern count in total number of patterns extracted and also the time spent on page[19].

3. Methodology

Web Log Data is huge and initial data cleaning is done in preprocessing stage to remove entries which are irrelevant for mining process. Users and sessions are identified and the sessions are reconstructed to form a User session matrix which consists of web pages as attributes and navigation patterns as rows. Analyzing user navigation patterns is an active research area to understand access patterns and usage trends. Efficiency of navigation patterns are calculated by using three parameters such as frequency, utility and downloads[20]. Feature Selection is an important preprocessing task where only relevant attributes are selected since some pages are repeatedly browsed by users and some pages are sparsely browsed. Clustering and classification are two major techniques of mining research for grouping the existing users and to classify a new user by analyzing their browsing patterns.

Each of these tasks is interrelated and the output of one phase is used as an input for another phase. A number of techniques have been developed to carry out the above mentioned tasks in various applications. Based on the applications, the algorithms and techniques used for clustering, classification and feature selection vary correspondingly. The input for the proposed method is the user session matrix in which navigation patterns are the rows and features are columns. The overall method is diagrammatically represented in (Figure 1).

3.1 Feature Selection

Feature selection method is used to identify important features in the dataset, and discard other irrelevant and redundant features. It reduces the dimensionality of the data, to hold the possibility of more effective and rapid operation of data mining algorithms and also the accuracy of future classification is improved. There are three types of feature selection algorithms in machine learning literature as filter methods, wrapper methods, and embedded methods. Filter methods are selected for the proposed work due its various advantages than the remaining two as simple, fast computation, not specific to any classifier, scales easily to very high dimensional datasets.

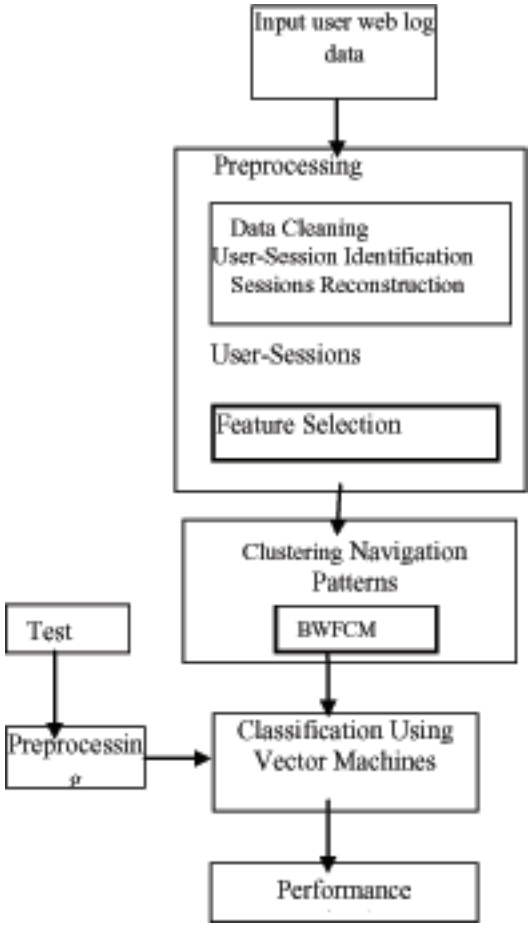

Figure 1. Framework of web log analysis.

3.1.1 Feature Selection using Independent Component Analysis

Independent Component Analysis (ICA) is a rich statistical technique developed for digital signal processing applications for revealing hidden factors for measurement of signals. Due to its versatile application in different areas such as source separation, channel equalization, speech recognition and functional magnetic resonance imaging, face recognition, telecommunication, predicting stock market place and financial market data mining[21] ICA has gained attention in recent years. This method finds a linear transformation in which the extracted components are mutually independent from each other and it focuses only on the statistical independence of features in the high dimensional data. It also acts as a powerful tool for analyzing text document data if the documents are presented in a suitable numerical form and to represent word histograms[22].

To select the best attributes also termed as features in user-session matrix ICA is used. Each and every feature is normalized by calculating mean and standard deviation. Absolute mean is calculated for all rows. Independent matrix is created by comparing and shrinking the values. Finally the independent matrix is multiplied with original matrix and the mean is calculated for all attributes. The mean is compared with a threshold and attributes which are less than threshold are selected.

3.1.2 Feature Extraction using ICA

1. Read session matrix with 'N' columns which are features and 'T' rows of navigation patterns and is denoted as x= $[x_1, \ldots, x_N]^T$ and 'c' classes.
2. Calculate mean 'm_i' and standard deviation 'σ_i' of each feature f_i.
3. By using formula $(f_i - m_i)/2\sigma_i$ where m_i and σ_i are the mean and the standard deviation of f_i, normalization of each feature is calculated.
4. Absolute mean for each user navigation patterns T, independent row vector W_i of W is computed as

$$a_i = \frac{1}{N+1} \sum_{j=1}^{N+1} |w_{ij}|$$

5. For all w_{ij} in W, if $|w_{ij}| < a_i$, then shrink $|w_{ij}|$ to zero to form new weight matrix W'.
6. Mutiply W' and the original user navigation pattern data X to construct a new matrix FS.

FS= {fi = W'iX, i∈ 1 • • • N+1}

7. Delete f_i, if the corresponding weight for class w_{ic} and w_{ij} is zero for all j ∈ 1 . . . N.
8. Resulting FS contains extracted features for user navigation patterns.

Feature Selection process reduced the size of data set considerably. It improves the accuracy of clustering process due to low dimensionality and enhances the classification process with a decrease in time complexity of overall mining process.

3.2 Clustering User Sessions into Groups

The Fuzzy C-Means (FCM) clustering algorithm is the best known and most powerful methods used in cluster analysis. Data elements belong to more than one cluster, and associated with each element in a set of membership levels. The levels indicate the strength of the association between that data element and a particular cluster. Some elements are in the center and some elements are in edge. The points on the edge of a cluster may be in the cluster to a lesser degree than the points in the center of cluster. The procedure of FCM is intended mainly to minimize the objective function of FCM with centroid values calculated at each iteration. The objective function is defined as follows.

$$J_{FCM} = \sum_{K=1}^{N} \sum_{i=1}^{C} (u_{ik})^a \, d^2 (X_k, V_1)$$

where 'd' is the distance between centroid and pattern, 'u_{ik}' is the membership value, 'x_k' is the pattern and 'v_i' is the centroid.

3.2.1 Bolzana_Weierstrass Theorem

The Bolzano–Weierstrass theorem deals with the results in convergence of a finite-dimensional Euclidean space R^n. The theorem is named after Bernard Bolzano and Karl Weierstrass[23]. It states that "Every bounded sequence has a convergent subsequence".

This is a nested interval theorem to imply the intersection of all the intervals $[a_n, b_n]$ is a single point 'w'. The theorem is proved with a sequence of 'n' numbers. Let $\{w_n\}$ be a bounded sequence. Then, there exists an interval $[a_n, b_n]$ such that $a_1 \le w_n \le b_n$ for all n. A sequence of intervals $\{[a_n, b_n]\}$ can be obtained by mathematical induction as follows.

1. For each n, $[a_n, b_n]$ contains infinitely many terms of $\{w_n\}$
2. For each n, $[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n]$
3. For each n, $b_n + 1 - a_n + 1 = \frac{1}{2}(b_n - a_n)$

3.2.2 Bolzano-Weierstrass FCM

The proposed method is carried out in two phases. In the first phase traditional FCM is implemented and clusters 'c' are formed and validation of clusters followed by reclustering is done in the second phase.

Step 1: The algorithm is carried out by clustering iteratively where an optimal c partitions are created by minimizing the weighted within group sum of squared error objective function J_{FCM}.

$$J_{FCM} = \sum_{k=1}^{n}\sum_{i=1}^{c}\left(u_{ik}\right)^{a} d^{2}\left(x_{k}, v_{1}\right)$$

Where $x_1, x_2, \ldots x_n \in \{user\ navigation\ patterns\}$, 'c' denotes the number of clusters with $1 \le c < n, = u_{ik}$ is the degree of membership of x_k in the ith cluster, 'q' is the weighting exponent on each fuzzy membership, 'v_{ik}' is the centroid value of the cluster i, $d^2_{k,t}$ is the Euclidean distance from user navigation pattern x_k to the cluster center v_i.

The number of clusters 'c', weighting exponent 'q' and an error value 'ε' are initialized. A fuzzy membership matrix 'U' is created for the input matrix whose values are between 0 to 1 with the following constraints.

$$\sum_{k=1}^{c} u_{k,i} = 1 \forall k$$

$$\sum_{k=1}^{c} u_{k,i} > 0 \forall k$$

Centroids are estimated with membership values for all clusters using the formula

$$v_k^{(b)} = \frac{\sum_{k=1}^{c}\left(u_{ik}^{(b)}\right)^{a} X_{ik}}{\sum_{k=1}^{c}\left(u_{ik}^{(b)}\right)^{a}}$$

The Euclidean distance is calculated between the centroids and navigation patterns.

$$d_{k,i} = \sqrt{\sum_{j=1}^{n}\left(V_{k,j} - X_{i,j}\right)^{2}}$$

The fuzzy matrix is updated with new membership values given as

$$u_{ik}^{(b+1)} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{k,i}}{d_{k,j}}\right)^{\frac{2}{q-1}}}$$

The pattern with minimum distance with any centroid is computed and the matched patterns are removed from the input matrix. The process is an iterative one. The centroid values are updated again with new membership values. Euclidean distance is calculated again with the updated values.

$$I_k = \left\{ i \,|\, i \le c,\ d_{ik} = \left\| X_k - V_1 \right\| \right\}$$

$$\hat{I} = \{1, 2, \ldots, c\} - I_k$$

where I_k is the distance between pattern and centroid and $\check{I}_k$ are the patterns remaining after first pattern is removed. The process is repeated until the difference between two consecutive fuzzy membership matrix are not more than the error value 'ε'

After desired results are obtained defuzzification is carried out. Fuzziness helps us to evaluate the rules, but the final output of a fuzzy system has to be a crisp number. Many methods are available for defuzzification. The method chosen for proposed work is Lambda-cut method which produces crisp sets called Lambda-cut sets. The values are formed by the following formula

$$A_\lambda = \{x / \mu_A(x) \ge \lambda\}$$

The resultant clusters are validated and optimized by Bolzano_Weierstraass theorem in the second step.

Step 2: Bolzano-Weierstrass is a theorem about sequences of real numbers[21]. As per theorem a set C_1 which contains the following clusters, ($C_2, \ldots . C_n$) the number of clusters results from FCM and it is bounded with Euclidean distance space $\mathbf{R}^p$ and without loss of patterns in the cluster. As per theorem the cluster set is considered as a sequence of real numbers (x_n) and it is bounded. Then divide the set of cluster ($C_2, \ldots . C_n$) into two sets I_1 & I_2. The process starts its validation by computing distances between two clusters in a subset. Perform the clustering between the same sets of clusters within I_1 and I_2. Initially choose $x_i \in C_i$ where 'x_i' is a pattern and 'C_i' is a cluster in a subset either in I_1 or I_2 is. Then define a set A = {x_i} which have infinitely many user navigation patterns as per theory. If the user navigation pattern set is bounded $A \subseteq C_2$ and $n_2, \ldots . n_n$ in each cluster, the numbers of patterns are compared based on a threshold. To perform cluster analysis first check the clusters that contains at least minimum number of patterns. Consider another cluster B = {x_j} is finite where x_j, the patterns in set B or another cluster are in the same subset. Let the distance between the two patterns is measured based on the parameter

$$r_i = \left\| X_j - X_i \right\|, X_j \in B, X_1 \in A$$

If 'r_i' is less than the minimum distance threshold value MDT then combine x_j pattern to cluster A. MDT should be less than the distance measure in FCM. If the above condition is not satisfied within the set I_1, then select clusters in I_2 and perform the same process.

The procedure for BWFCM algorithm is given as follows.

1. The values 'c', 'q', 'ε' are initialized. Initialize the values c, q, and ε. Initialize fuzzy membership matrix with the following constraints.

$$\sum_{k=1}^{c} u_{k,i} = 1 \forall k$$

$$\sum_{k=1}^{c} u_{k,i} = 0 \forall k$$

Where $u_{k,i} = u_k(X_i), 1 \le c$ and $1 \le i \le n$. Let $u_{k,i}$ satisfy the above conditions represented by a c · n matrix $u = [u_{k,i}]$.

2. Set the iteration counter b = 0.
3. Calculate the c cluster centres $\{v_k^b\}$ with $U^{(b)}$.

$$v_k^{(b)} = \frac{\sum_{k=1}^{c} \left(u_{ik}^{(b)}\right)^a X_{ik}}{\sum_{k=1}^{c} \left(u_{ik}^{(b)}\right)^a}$$

where the exponent 'q'is the degree of fuzziness associated with the partition matrix.

4. Calculate Euclidean distance by using formula

$$d_{k,i} = \sqrt{\sum_{j=1}^{n} \left(V_{k,j} - X_{i,j}\right)^2}$$

5. Calculate the membership $U^{(a+1)}$.
6. For, $k = 1$ *to c,* calculate the following:

$$I_k = \left\{ i \middle| 1 \le i \le c,\ d_{ik} = \left\| X_k - V_1 \right\| \right\}$$

$$\hat{I}_k = \{1, 2, \ldots, c\} - I_k$$

for the k^{th} column of the matrix.

Else

$u_{ik}^{(b+1)} = 0$ for all If $l_k \neq \emptyset$ compute new fuzzy partition matrix U with membership values

$$u_{ik}^{(b+1)} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{k,i}}{d_{k,j}}\right)^{\frac{2}{q-1}}}$$

$i \in \hat{I}_k$ and, $\sum_{i \in \hat{I}} u_{ik}^{(b+1)} = 1$

7. If $\left\| U^{(b)} - U^{(b+1)} \right\| > \varepsilon$, increment $b = b + 1$ go to step 3.
8. Combine fuzzy sets and defuzzify each value using

$$A_\lambda = \{x / \mu_A(x) \ge \lambda\}$$

9. Repeat for all $X_n \in C_n$
10. Divide the cluster results into two set such as I_1 & I_2,
11. Select different clusters $X_i \in A$, $X_j \in B$ in same set I_1
12. Calculate distance between the patterns in A and B

$$r_i = \left\| X_j - X_i \right\|, X_i, X_j, \in A, X_j \in B$$

13. If r_i < MDT then A = A + X_j, else compare with remaining clusters in the same set.
14. If no cluster matches in the same subset select clusters from another subset and Go to step 8.
15. Stop the process.

3.2.3 *Flowchart for BWFCM Process*

The pictorial representation of step by step procedure is shown in (Figure 2).

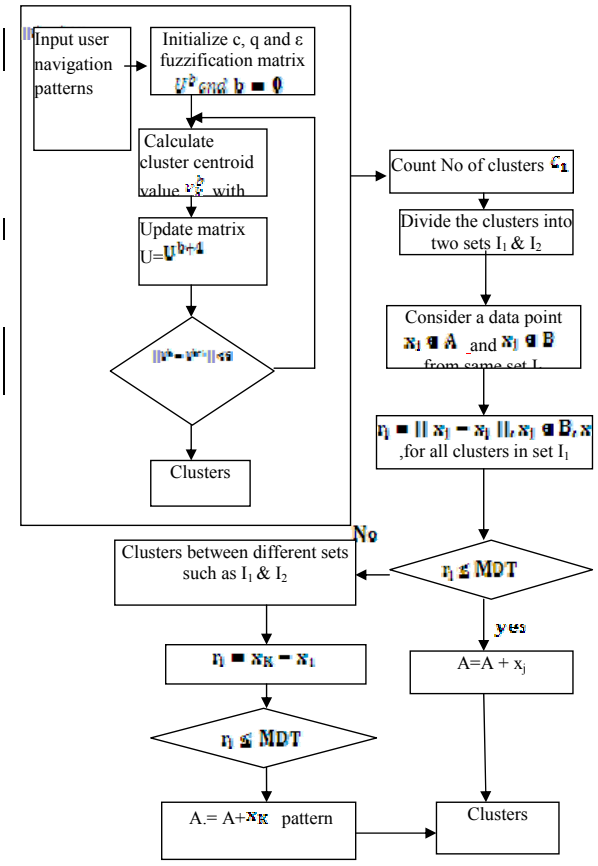

Figure 2. BWFCM process.

4. Experimental Results

The proposed methodology is evaluated in this section. The results of different stages are discussed. The method is implemented by using MATLAB.

4.1 Web Log Data

User's requests are accumulated in web server and it is an automatic process performed by the web server. The data sets used in this work are three real time datasets. All three sets were collected from web logs of IIS web servers installed for an e-commerce web server, academic institution web server and a research journal web server for a period of 3 months from October to December, 2013. Due to privacy concerns the details of log files are not exposed. To prove the applicability of proposed work in different domains, the data sets were collected from three different web logs. The existing and proposed methods were implemented in the web log sets.

4.2 Preprocessing of Log Data

The logs collected had 30,232 entries from e-commerce website, 56596 entries from academic institution website and 6697 entries from the research journal's website. Preprocessing was the first step in this process to enhance accuracy, speed in other mining process. The steps carried out in preprocessing are discussed below. Data cleaning phase is performed and irrelevant entries are removed. Log entries requested for graphics and video formats such as gif, JPEG, etc., are removed. Sessions are identified by Time oriented method with 30 minutes. Totally 999 sessions in e-commerce log file, there are 1400 sessions in the institutions' log, and in the research journal's log there are 180 sessions identified. The sessions are restructured as rows in a navigation pattern matrix and web pages navigated as attributes or features. When a user visit a page in the session an entry is posted and incremented for each revisit in the page.

ICA is applied in the user session matrix to select relevant and important features. The result of each and every step in preprocessing is tabulated below (Table 1).

4.3 Clustering Accuracy

Internal quality metrics are used to measure the similarity of cluster elements using some measure. It usually measures the intra-cluster homogeneity, the inter-clusters dissimilarities. To measure the clustering accuracy of proposed and the existing methods, the parameters such as Rand Index, Sum of Squared Error (SSE), F measure are calculated and compared.

Table 1. Preprocessing Results

Entries taken	E-commerce	Institution	Research
Number of samples	30,232	56596	6697
After cleaning	19239	51148	6005
Users Identified	356	890	65
Sessions Identified	990	1400	180
Features in data sets	14	13	7
No of features selected	6	8	4

4.3.1 Rand Index

The Rand index is a measure used to compare clusters. Let C_1 and C_2 are two clusters considered for evaluation. 'a' be the number of navigation patterns assigned to the cluster in C_1 and C_2. 'b' be the number of patterns assigned in the cluster C_1 and not in the cluster in C_2. 'c' be the number of patterns that are in the cluster C_2 but not in C_1 and 'd' be the number of patterns that are assigned to different clusters in C_1 and C_2. The values a and d are interpreted as agreements, and b and c as disagreements. The Rand index is defined as

$$\text{Rand Index} = \frac{a+d}{a+b+c+d}$$

The resultant value of Rand index lies between 0 and 1. When the clustering process is carried out perfectly, the Rand index is 1. The clustering accuracy is high and the rand index value is nearly equal to one or else the accuracy of clustering results is less. It shows proposed BWFCM have higher rand index rate than FCM. The results of clustering methods for Rand Index are tabulated in Table 2 and plotted in Figure 3.

4.3.2 Sum of Squared Error (SSE)

SSE is the simplest and most widely used measure for clustering. It is calculated as:

Table 2. Comparison of rand index

Dataset	Rand index comparison	
	FCM	BWFCM
E-Commerce	0.74	0.92
Academic	0.75	0.94
Researchers	0.72	0.95

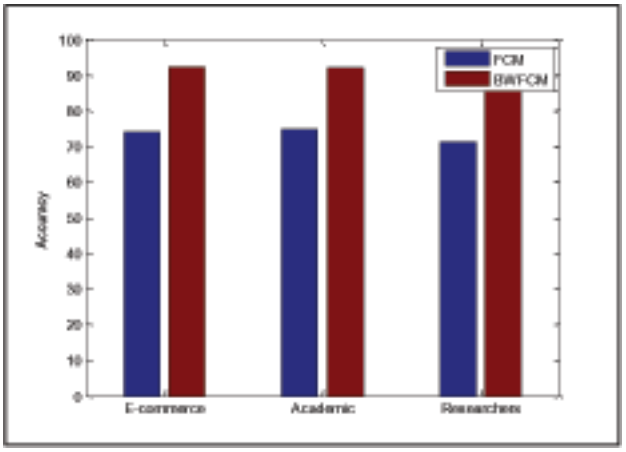

Figure 3. Comparison of rand index.

$$SSE = \sum_{k=1}^{K} \sum_{\forall X_i \in C_k} \left\| X_i - \mu \right\|^2$$

where C_k is the set of patterns in cluster k, μ_k is the vector mean of cluster k. The components of μ_k are calculated as:

$$\mu_{k,j} = \frac{1}{N_k} \sum_{\forall X_i \in C_k} X_{1-j}$$

where N_k is the number of patterns belonging to cluster k. and calculated as $|C_k|$ The calculated values are tabulated in table 3.

From the above experiments the sum of squared error gives more valid results and simple. For web log clustering this is a most suitable one.

4.4 Impact of Feature Selection in Clustering

To understand the impact of the feature selection method, the data sets were implemented with the existing and proposed methods of feature selection. The parameter taken for comparison is F-measure. The method was implemented in all the three data sets after data cleaning, session construction step. Clustering was carried out twice with the proposed clustering algorithm in all the three preprocessed data sets. First time the BWFCM clustering is done once the user session matrix with navigation patterns had been created. Second time the method was applied in the reduced matrix after the features were selected with Independent Component Analysis. Once clusters were formed the F-measure was calculated for all new clusters which were formed.

Table 3. Comparison of SSE

	SSE Comparison	
Dataset	FCM	Proposed BWFCM
E-commerce	0.5771	0.3287
Academic	0.6113	0.3296
Researchers	0.6781	0.3779

It was observed that the clustering results are poor when applied in full data set with irrelevant features, and the performance was increased after relevant features were selected. The results are tabulated in table 4. Time taken for the process also reduces with feature selected datasets. Random samples are drawn and executed ten times and results are depicted as follows (Figure 4).

5. Conclusion

Analysis of user navigation patterns is a useful research area due to its applicability in various domains and in the present work it was carried out to know the users interests implicitly by creating sessions and these user sessions are

Table 4. F measure without feature selection and with feature selection

	F-Measure	
Dataset	BWFCM Without FS	BWFCM with ICA
E-Commerce	0.3226	0.4839
Academic	0.3346	0.4568
Researchers	0.3342	0.4495

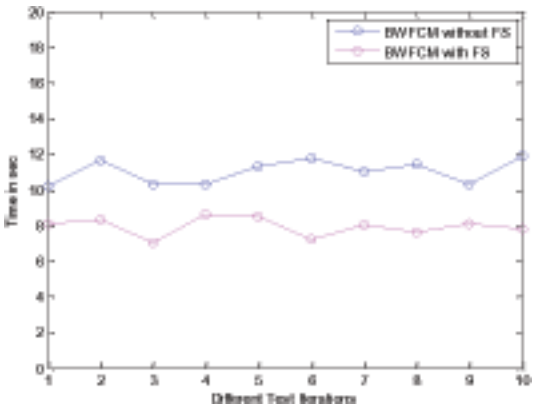

Figure 4. Comparison of resultant time taken before and after Feature Selection

matched with a new user's sessions. It is carried with three important techniques namely feature selection, clustering and classification. In this paper a novel clustering optimization method is implemented and results are evaluated. Impact of feature selection process in the clustering results is also evaluated. The result of proposed clustering method optimized the clusters formed by Fuzzy C Means to enhance classification process in the next phase of analysis of navigation patterns. With the emerging new applications over the internet many new types of web data such as email traffic, web blogs and wiki pages are available which produces a large amount of new knowledge resources where the system can be remodeled a little for its use.

6. References

1. DeMin D. Exploration on web usage mining and its application. International Workshop on Intelligent Systems and Applications; 2009: IEEE; 2009.
2. Hofsesang. Methodology for preprocessing and evaluating the time spent on web pages. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence; 2006.
3. Zheng L, Gui H, Li F. Optimized data preprocessing technology for web log mining. Proceedings of the International Conference on Computer Design and Applications (ICCDA) Vol. 1; IEEE; 2010.
4. Fodor IK. A survey of dimension reduction techniques. LLNL Technical Report. 2002. UCRL-ID-148494.
5. Choi SW, et al. Fault detection based on a maximum-likelihood Principal Component Analysis (PCA) mixture. Ind Eng Chem Res. 2005.
6. Rathipriya R, Thangavel K. Extraction of web usage profiles using simulated annealing based biclustering approach. 2014. arXiv preprint arXiv:1412.8099.
7. Yu Y-X, Wang X-W. Web Usage mining based on fuzzy clustering. International Forum on Information Technology and Applications; IEEE; 2009.
8. Yang M, Li H. User analysis based on fuzzy clustering. International Conference on Business Intelligence and Financial Engineering; IEEE; 2009.
9. Suresh K, et al. Improved FCM algorithm for Clustering on Web Usage Mining. 2011 International Conference on Computer and Management (CAMAN); IEEE; 2011.
10. Shi P. An efficient approach for clustering web access patterns from web logs. International Journal of Advanced Science and Technology. 2009; 5(1).
11. Vaishnavi A. Effective Web Personalization System using Modified Fuzzy Possibilistic C Means. Bonfring International Journal of Software Engineering and Soft Computing 1. 2011; Inaugural Special Issue.
12. Maji P, Pal SK. RFCM: A hybrid clustering algorithm using rough and fuzzy sets. Fundamenta Informaticae. 2007.
13. Vanisri D, Loganathan C. An efficient fuzzy possibilistic C-Means with penalized and compensated constraints. GJCST. 2011.
14. Joshi A, Krishnapuram R. On mining web access logs. Baltimore County Baltimore MD: Maryland University; 2000.
15. Nasraoui O, Gonzalez F, Dasgupta D. The fuzzy artificial immune system: Motivations, basic concepts, and application to clustering and web profiling. Proceedings of the World Congress on Computational Intelligence (WCCI) and IEEE International Conference on Fuzzy Systems; 2002.
16. Kanade PM, Hall LO. Fuzzy ants as a clustering concept. NAFIPS 2003. 22nd International Conference of the North American; Fuzzy Information Processing Society: 2003: IEEE; 2003.
17. Zhao Z, Bai K. Research on T-bridge algorithm of web session fuzzy clustering. 2008 IFIP International Conference on Network and Parallel Computing; 2008.
18. Revathy S, Parvaathavarthini B, Rajathi S. Futuristic validation method for rough fuzzy clustering. Indian Journal of Science and Technology. 2015 Jan.
19. Mehta P, Jadhav SB, Joshi RB. Web usage mining for discovery and evaluation of online navigation pattern prediction. Int J Comput Appl. 2014 Apr.
20. Grace, LKJ, Maheswari V. Efficiency calculation of mined web navigational patterns. Indian Journal of Science and Technology. 2014.
21. Hyvarinen JK, Oja E. Independent component analysis. New York: Wiley; 2001.
22. Kolenda LKH, Sigurdsson S. Independent components in text. In: Girolami M, editor. Advances in Independent Component Analysis. Springer-Verlag; 2000.
23. Ellermeyer, Bolzano-Weierstrass theorem, 2000.