Analysis for Identifying Best Fit Clustering Algorithm on Web Usage Data

Naveen Kumar Penki, M. Rekha Sundari, Y. Srinivas

Abstract— **Web Usage Mining (WUM) is the important application of data mining techniques to discover interesting usage patterns from web access log. Analysing the web usage logs helps understand the user's browsing behaviour. In our work we implement the K-means (KM), K-medoids and Fuzzy C-Means algorithms on web log data. These algorithms group the users with similar browsing behaviour into the same cluster. A single Algorithm may not be accurate for all types of data sets. We may not initially know which algorithm fits to our data set best. So we compared the clustering accuracy of these algorithms using different accuracy measures available and by analysis we conclude the algorithm that is best suited for web usage data**

Index Terms— **Clustering, K-means, K-medoids, FCM**

I. INTRODUCTION

Data Mining (DM) is the extraction of information from large amounts of data to view the hidden knowledge and facilitate the use of it to the real time applications. DM consists of algorithms for predictive and descriptive analysis of data. Some of the core Data Mining techniques used for these analysis are Clustering, Association Rule Mining, Classification and etc. Clustering is an effective technique for data analysis. The existing methods of clustering can be categorized into three: partitioning, hierarchical, grid-based and model-based methods. Partition clustering generates a division of the data such that the distance between the data points in the same cluster is minimized than to the data points in other clusters. Clustering is a technique to search invisible patterns that exists in datasets. It is a process of grouping data objects into disjoint clusters so that the data in each cluster are similar, yet different to the other clusters. Some of the clustering methods that minimize the clustering error are the k-means, k-medoids, and FCM algorithms. These are attractive, because of simplicity and speed in computing. It splits the input dataset into k clusters. Each cluster is represented by changing centroid (also called cluster Centre), starting from some initial values named seed-points.

This paper deals with comparison of partitional clustering algorithms on web usage data to cluster web users of similar browsing behaviour. Here we are going to implement K-means, k-medoids and FCM algorithms on a web user data set and analyse which algorithm is best suited for clustering web usage data set.

II. RELATED WORK

JinHua Xu et.al [1] Explore the problem of user clustering based on vector matrix and K-Means algorithm. Proposed a technique to cluster web usage data using simple algorithm K-Means for clustering. This method generates a vector matrix which is used to describe the relationship between web pages and different users who accessed at specific session. Cosine similarity is used as a similarity measure. Then the K-Means algorithm is applied to cluster the data set. The result show that proposed algorithm is feasible and have scalability

Noor Kamal Kaur et.al [2] proposed the partitioning method that can be performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point. This forms the basis idea behind k-medoids method. The basic strength of k-medoids clustering algorithms is to find k clusters in n instances by first arbitrarily finding a representative instance for each cluster. Each remaining instance is clustered with the medoid to which it is the most similar. The k-medoids method uses representative instances as reference points instead of taking the mean value of the instances in each cluster. The algorithm takes the input criterion k, the number of clusters to be divided among a set of n instances. In this paper it is concluded that k-medoids showed better performance than k-means according to the no of links and execution time taken by both the algorithms to form the clusters.

K. Suresh et.al [3] proposed a method, information entropy to initialize cluster canters and introduced weighting parameters to adjust the index of cluster canters as the clustering method is very sensitive to initial canter values. First of all here the information entropy is used to initialize the cluster canters to determine the number of cluster canters. It finally divides the large chumps into clusters. Then merge various small clusters according to the merger of the conditions, so that you can solve the irregular datasets clustering. Here web usage data is clustered which is useful in finding the user access patterns and the order of visits of the hyperlinks of the each user. It improves the efficiency of Fuzzy C Means clustering algorithm.

Manuscript received June 07, 2017

Naveen Kumar Penki, has completed his bachelors of Technology in Computer Science & Engineering at GITAM University, Visakhapatnam. His research areas includes Data Mining, Cryptography and Text Analytics

M. Rekha Sundari, M completed her M.Tech from GITAM University, Visakhapatnam and Ph.D. from JNT University Kakinada. She is presently working as Assistant Professor in Department of Computer science and Engineering, GITAM University, Visakhapatnam. Her research area includes Data Mining and Image Processing

Y. Srinivas, Srinivas Y is currently working as a Professor, in Department of Information Technology, GITAM University, Visakhapatnam. His research area includes Image Processing, Data Mining, and Software Engineering

S. Revathi et.al [4] proposed a theory for comparison of various clustering algorithms. A comparative study of clustering algorithms across two different data sets is performed using Simple K-Means Clustering, Efficient K-Means Clustering, Farthest First Clustering, Density Based Clustering algorithms. The experimental results of various clustering algorithms to are elucidated as a graph. Thus it is concluded as the time taken to form the clusters increases as the number of clusters increases. The Farthest First Clustering algorithm takes very few seconds to cluster the data items whereas the Simple K-Means Clustering algorithms takes the longest time to perform the clustering.

K. Santhisree et.al [5] presented new Rough set DBSCAN clustering algorithm which identifies the behaviour of the users page visits, order of occurrences of the visits. The experimental results on msnbc.com which is useful in finding the user access patterns and the order of visits of the hyperlinks of the each user and the inter cluster similarity among the clusters. In rough set agglomerative clustering the elements can be presented in more than one cluster where in rough set DBSCAN, the elements will not occur in other clusters. The Rough set DBSCAN clustering algorithm is efficient when compared to the rough set agglomerative clustering.

III. METHODOLOGY

Here we have taken a web user data set from msnbc.com which is sequential and unstructured initially.It consists of web pages and user page visit data.As clustering cannot be applied on unstructered data we need to process the data which gives a vector matrix of users and webpages. This data set consists of 349 users and 17 different web pages that are browsed by users. This processed dataset and the k-value(No of clusters:3)is given as input to the three algorithms.

Algorithms Implemented
1.K-Means
2.K-medoids
3.Fuzzy C-Means

These three algorithms results in clusters of similar user behaviour (page visits) as output.Here minkowski measuring technique is used to find the distance between the users or objects and Silhouette measuring technique to find the accuracy of clusters formed by the algorithms.Based on the silhouette values and graphs constructed for user page visits the algorithm that clusteres the users in a better way is concluded.

SYSTEM DESIGN:

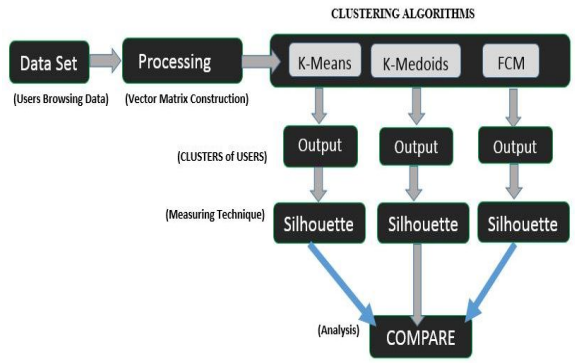

Fig 1: System Model

INPUT:

No of Users : 349
No of Pages : 17
No of Clusters : 3
Initial Centroids : {61,142,269}

OUTPUT(Graph Constructed from data) :

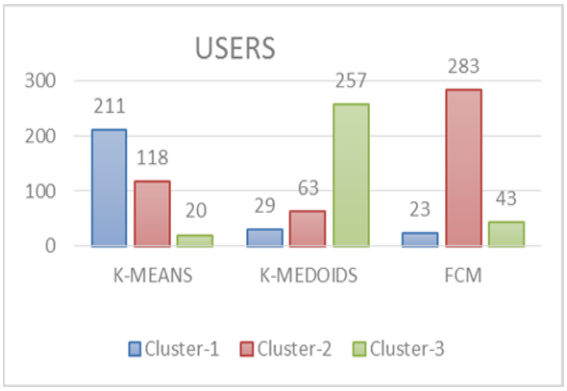

Fig 2: Clusters Formed by algorithms

These clusters are analysed based on the total number of users within a cluster who have visited a particular web page.
Thus we can generate graphs for the clusters formed by the three algorithms for the 17 web pages are as below

By analysing the graphs we can have a clear picture of clusters formed based on the users browsing behaviour i.e. users with less number of page visits are grouped in one cluster, moderately visited in one cluster and regularly visited are grouped into other.

K-Means Algorithm:

The algorithm is composed of the following steps:
1. Select k points in the space indicated by the instances that are to be clustered. These points are initial group centroids.
2. Assign each instance to the group with closest centroid.
3. When all instances have been assigned, recalculate the positions of the new centroids.
4. Repeat Step 2 and 3 until the centroids no longer converge.

This algorithm calculates mean of clusters to find new centroids.

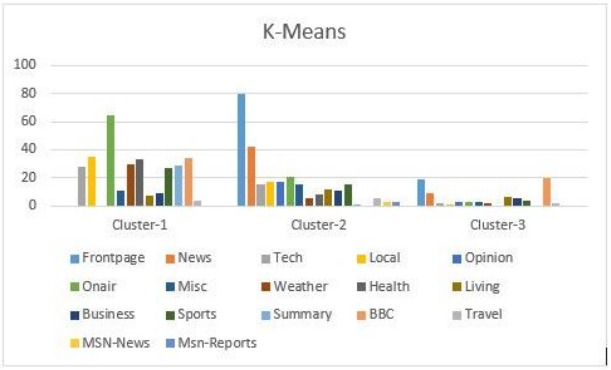

Fig 3: Clusters Formed by K-Means algorithm

K-Medoids Algorithm:

Method: Randomly choose k instances in Dataset D as the initial representative points.

Assign repeatedly each of the remaining instances to the cluster with the nearest medoid randomly select a non medoid object Orandom.

Compute the total points S of swapping object Oj with Orandom if S < 0 then swap Oj with Orandom to form the new set of k medoid until no change.

This algorithm uses cost function to calculate new centroids.

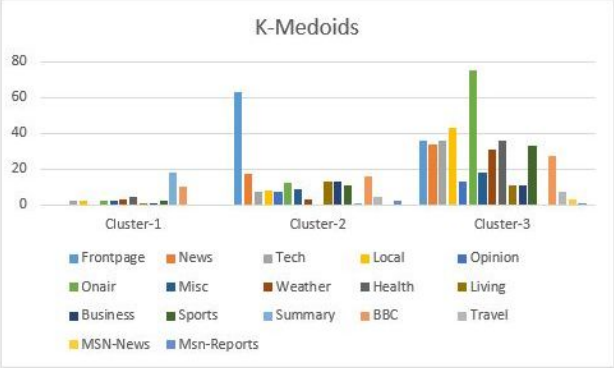

Fig 4: Clusters formed by K-Medoids.

Fuzzy C-Means Algorithm (FCM):

Initialize number of clusters, Cj(cluster centers),
□□(threshold value)
Repeat
For i=1 to n: update μj(Xi)
For k=1 to p ;
Sum= 0, Count=0
Fori=1 to n:
If μ (Xi) is maximum in Ck
Then If μ (Xi)>= □ Sum=sum+Xi
Count= count+1
Ck=sum/count
Until Cj estimate stabilize.

This algorithm calculates fuzzification parameter to calculate new centroids.

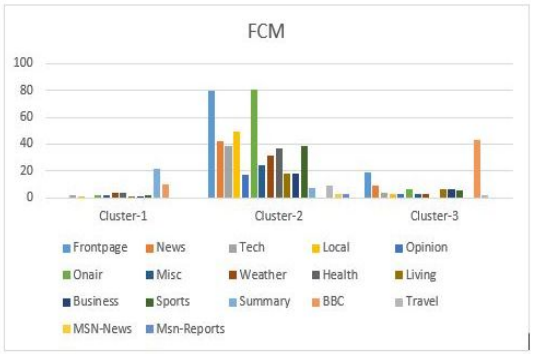

Fig 5: Clusters formed by FCM(Fuzzy C Means)

Silhouette Co-efficient:

An efficient measuring technique, the silhouette width, that shows good performance in experiments, was introduced by Kaufman and Rousseau. The concept of silhouette width involves the difference between the within-cluster accurateness and separation from the rest. Specifically, the silhouette width *s(i)* for entity *i*is defined as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Where *a (i)* is the average distance between *i* and all other instances of the cluster to which *I* belongs and *b (i)* is the minimum of the average distances between *i* and all the instances in each other cluster. The silhouette width values lie in the range from—1 to 1. If the silhouette width value for an instance is zero, it means that that the entity could be assigned to another cluster as well. If the silhouette width value is close to—1, it means that the instance is wrongly clustered. If all the silhouette width values are close to 1, it means that the set I is well clustered.

SILHOUETTE VALUES:
Number of Clusters: 3

K-Means : 1.1436673
K-Medoids : 1.3887466
FCM : 1.6466541

From the above silhouette values of each clustering algorithm, it is observed that the silhouette value of the FCM algorithm is higher than the other two algorithms which indicates that, the instances are clustered more accurately and efficiently in FCM clustering algorithm than other two algorithms.

IV. CONCLUSION

In this paper we have implemented K-means, K-medoids and FCM clustering algorithms on a web usage data set taken from msnbc.com from which we have analyzed that K-means algorithm fails to generate clusters if the data set is sparse and the silhouette values of FCM are better compared to K-Means and K-medoids algorithms and the graphs of clusters of users browsing behavior in FCM are accurate than the remaining algorithms. Hence Fuzzy C-means clustering algorithm suits better for clustering this dataset.

V. REFERENCES

[1] Jin Hua Xu, Hong Liu, "Web User Clustering Analysis based on KMeans Algorithm", International Conference on Information, networking and Automation (ICINA), IEEE, 2010.

[2] Noor Kamal Kaur, Usvir Kaur, Dr.Dheerendra Singh "K-Medoid Clustering Algorithm- A Review" International Journal of Computer Application and Technology (IJCAT) Volume 1, Issue 1, April 2014

[3] K.Suresh, R.MadanaMohana, A.RamaMohanReddy, A.Subrmanyam "Improved FCM algorithm for Clustering on Web Usage Mining" IEEE, 2011.

[4] S. Revathi, Dr.T.Nalini "Performance Comparison of Various Clustering Algorithm" International Journal of

Advanced Research in Computer Science and Software Engineering Volume 3, Issue 2, February 2014.

[5] K.Santhisree, Dr A. Damodaram, S.Appaji D.NagarjunaDevi, "Web Usage Data Clustering using Dbscan algorithm and Set similarities" International Conference on Data Storage and Data Engineering, IEEE, 2010.

AUTHORS:

Naveen Kumar Penki, has completed his bachelors of Technology in Computer Science & Engineering at GITAM University, Visakhapatnam. His research areas includes Data Mining, Cryptography and Text Analytics

M. Rekha Sundari, M completed her M.Tech from GITAM University, Visakhapatnam and Ph.D. from JNT University Kakinada. She is presently working as Assistant Professor in Department of Computer science and Engineering, GITAM University, Visakhapatnam. Her research area includes Data Mining and Image Processing

Y. Srinivas, Srinivas Y is currently working as a Professor, in Department of Information Technology, GITAM University, Visakhapatnam. His research area includes Image Processing, Data Mining, and Software Engineering