SSM-DBSCANand SSM-OPTICS : Incorporating a new similarity measure for Density based Clustering of Web usage data.

Ms K.Santhisree [1], Dr. A Damodaram [2]
[1,2]Dept. of Computer science, Jawaharlal Nehru Technology University (JNTUH), Hyderabad, India
[1] santhisree.k@jntuh.ac.in , [2]damodaram@rediffmail.com

Abstract: **Clustering web sessions is to group web sessions based on similarity and consists of minimizing the intra-group similarity and maximizing the inter-group similarity. Here in this paper we developed a new similarity measure named SSM(Sequence Similarity Measure) and enhanced an existing DBSCAN and OPTICS clustering techniques namely SSM-DBSCAN, and SSM-OPTICS for clustering web sessions for web personalization. Then we adopted various similarity measures like Euclidean distance, Jaccard, Cosine and Fuzzy similarity measures to measure the similarity of web sessions using sequence alignment to determine learning behaviors of web usage data. This new measure has significant results when comparing similarities between web sessions with other previous measures. We performed a variety of experiments in the context of density based clustering, using existing DBSCAN and OPTICS and developed SSM-DBSCAN and SSM-OPTICS based on sequence alignment to measure similarities between web sessions where sessions are chronologically ordered sequences of page visits. Finally the time and the memory required to perform clustering using SSM is less when compared to other similarity measures.**

Keywords: **clustering, similarity, SSM, SSM-DBSCAN,SSM-OPTICS sequential dataset; similarity measures, Intra cluster, Inter cluster.**

I. INTRODUCTION

Clustering is an initial and fundamental step in data analysis. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). Cluster analysis is the organization of a collection of patterns into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other and dissimilar when compared to a pattern belonging to a other cluster. Clustering plays an important role in data mining applications such as scientific ,data exploration, information retrieval and text mining. Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification. Broadly speaking, clustering algorithms can be divided into 4 types – Partitioned ,Hierarchical, Density based and Grid based methods. Partitioning algorithms are K-means and K-medoids. Hierarchical are Agglomerative and Divisive , density base are Dbscan,Optics and Denclue.

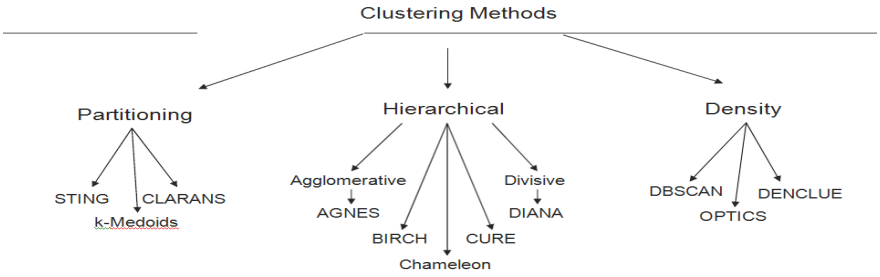

Fig 2.1:Types of Clustering Techniques

The structure of the paper is as follows, In unit-I we introduced some preliminaries on data mining, and reviewed the clustering algorithms. Then in Unit-II we discussed about work related to the density clustering. In Unit-III, density based algorithms were discussed. Then in Unit –IV, the description about

the similarity measures is discussed. In unit V, We introduced a a new similarity measure for clustering , which finds the distance between two sequences, and also developed a novel enhanced density based clustering algorithm incorporating the new similarity measure(SSM) into it namely SSM-DBSCAN and SSM-OPTICS. We carried out experimental comparison of several similarity measures like Euclidean, Cosine and Jaccard and Fuzzy similarity and with new similarity measure(SSM). Then conclusions in chapter VIII and finally the references.

II.RELATED WORK

More and more researchers focus on Web usage mining for the past recent years [3,4,5]. However, the topic of clustering web sessions has recently become popular in the field of practical application of clustering techniques. In [1] Anil K.J worked out on different algorithms for data and it performance. Few researchers in the past applied OPTICS clustering with noise on the different datasets and analyze on various ways. Ester .M.et.al[6] introduced about the various issues that related to the Dbscan for discovering clusters in large databases with noise. Cao.F, Estery .M, Qian .W [5],worked on Density based clustering over an evolving data stream with noise. "M Ankerst, M. Breunig, H.Kriegel, J.Sander" [11] introduce OPTICS algorithm on density based clustering structure. In [2] the authors described the various ways of scaling the Dbscan algorithm in the application of spatial database. Many researchers carried out their works on web usage clustering using Density based algorithms. The Density-based notion is a common approach for clustering. Density-based clustering algorithms are based on the idea that objects which form a dense regions should be grouped together into one cluster. They use a fixed threshold value to determine dense regions. Mobasher[15] used the Cosine coefficient and a threshold of 0.5 to cluster on a web log. Banerjee and Ghosh[9] introduced a new method for measuring similarity between web sessions. The longest common sub-sequence between two sessions is first found through dynamic programming, then the similarity between two sessions is defined through their relative time spent on the longest common sub-sequences.

III.DENSITY CLUSTERING ALGORITHMS

Density based clustering is to discovers clusters of arbitrary shape in spatial databases with noise . It defines a cluster as a maximal set of density connected points. The DBSCAN algorithm works on two factors on Density reach ability and ε -neighborhood. The problem is to cluster data which can be with or without noise. DBSCAN requires two parameters: ε (Eps) and the minimum number of points required to form a cluster (minPts). It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. DBSCAN's definition of a cluster is based on the notion of density reachability. Basically, a point *q* is directly density-reachable from a point *p* if it is not farther away than a given distance ε (i.e., is part of its ε-neighborhood), and if *p* is surrounded by sufficiently many points such that one may consider *p* and *q* be part of a cluster. *q* is called density-reachable from *p* if there is a sequence of p 1, p 2…. p_n of points with and $p_1 = p$ and $p_n = q$ where each p_{i+1} is directly density-reachable from p_i. Note that the relation of density-reachable is not symmetric, so the notion of density-connected is introduced. Two points *p* and *q* are density-connected if their is a point *o* such that *o* and *p* as well as *o* and *q* are density-reachable.

Then OPTICS ,the other clustering algorithm was designed by Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel and Jörg Sander. Its basic idea is similar to DBSCAN to specify clusters and noises of a special database. OPTICS defines a cluster with the base of density. Now we have two parameters first is *"MinPts"* and second is *"Eps"*,Where *"MinPts"* is the minimum points around "p" (is a node) and *"Eps"* is the maximum value for radius around the *"p"*.In this algorithm point "p" is a cluster if it contains minimum points (call *"MinPts"*) that are not farther than the defined distance (call *"Eps"*). If number of points with less distance of than *"Eps"* to *"p"* is more than *"MinPts"* then *"p"* is a cluster. If a point is a part of a cluster its name is e-neighborhood. OPTICS sees points that are part of a more densely packed cluster, so each point is imputed a core distance that basically describes the distance to its *"MinPts"* and finally if a point is neither a cluster and also is nor a part of any cluster then that point's name is Noise. OPTICS addresses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of varying density. In order to do so, the points of the database are (linearly) ordered such that points which are spatially closest become neighbors in the ordering. Additionally, a special distance is stored for each point that represents the density that needs to be accepted for a cluster in order to have both points belong to the same cluster. This is represented as a dendrogram.

IV. DISTANCE MEASURES

One of the important steps on clustering is distance measures. Distance measures, will find how two points in a cluster are similar.

Euclidean distance measure:The **Euclidean distance** between sequences p and **q** is defined as

$$d(p,q) = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}. \quad \text{.. (1)}$$

Jaccard similarity measure: The Jaccard similarity between sequences A and B, is defined as the size of the intersection divided by the size of the union of the two sequences:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \quad \text{.. (2)}$$

Cosine Similarity Measure: Cosine similarity is a measure of similarity between two vectors by finding the cosine of the angle between them.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (3)$$

Fuzzy Dissimilarity: Given two ordered fuzzy sets s1= (si1, si2, si3... sin) and s2= (sj1, sj2, sj3... sjn), is defined as

$$\text{Sim}(S1, S2) = \frac{(S1 \cap S2)}{(S1 \cup S2)} \quad (4)$$

V. PROPOSED APPROACH

A. A Novel Similarity Measure for Sequential Data.

A new similarity measure SSM(Sequence Similarity Measure) was devised for clustering of sequential data. This similarity measure(SSM) contains three parameters, one the Jaccard similarity (sequence similarity) and the second parameter the frequency count of the two sequences, and third the total length of the common sub- sequence common to both the sequences respectively. In the earlier distance measures, we have not considered the order of the sequence and frequency of the repetition of the web pages.

SSM between two sequences S1 and S2 is defined as

$$SSM = sim(S_1,S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2} * FC(S_1,S_2) + \frac{LLCS(S_1,S_2)}{\sqrt{\sum_{i=1}^{n}(S_{1_i})^2} \times \sqrt{\sum_{i=1}^{n}(S_{2_i})^2}} \quad \text{----(5)}$$

Where FC is the frequency count of the web pages ,LLCS is the length of the longest common sub sequence.

B. (SSM-DBSCAN) -An Enhanced density - based clustering algorithm

In order to cluster web pages we adopted existing DBSCAN and OPTICS clustering algorithms with three different distance/similarity measures. Consider a data set D ={T_1,T_2,T_3, …, T_N) with *N* web users surfing from a set of *m* webpage list {u_1,u_2,u_3,…,u_m}. Each user web page visit is represented as T_i = <u_1,u_2,u_3,…, u_k> and is called user session. Our objective is to segment the dataset D into different segments such that user sessions are clustered based on similarity. To determine whether a web user sessions are similar enough to be considered as a cluster we need a distance/similarity measure that determines how far/similar two web user sessions from each other. The most commonly used distance/similarity measures are, Euclidean and Jaccard, Cosine similarity measures. None of these measures captures the sequential nature of data as well as content

information of web sessions. We incorporated a new distance measure into this existing DBSCAN and OPTICS clustering techniques and a set of clusters are generated. Then for each cluster a cluster representative is selected. Then the inter cluster and the intra cluster similarities are computed .The objective of the algorithm is to find the clusters such that the Intra cluster similarity is minimum and the Inter cluster similarity is maximum. The steps to cluster web usage data can be summarized as follows

Algorithm:

Inputs : D = {T_1, T_2, …….Tn} is the dataset consisting of web user transactions where each T_i ={URLi1,URLi2, …..URLim} //URLs visited by user i. Eps is the radius , MinPts is the Minimum number of neighborhood points .

P €[0, 1]

Output:

C :Cluster group

Method:

STEP 1 : For each pair (T_i, T_j) € D,

STEP 2: Compute S(T_i, T_j) using a new SSM

from (equation 6).

C = NULL

Noise = NULL

STEP 3: For each T_i€ D

Mark T_i as visited

N = {T_k | T_k € D and S(T_i, T_k)≤ Eps}

If |N| < Minpts then

Noise = Noise U {T_i }

Else

Mark T_i as Visited

END

C. SSM-OPTICS: An Enhanced OPTICS clustering Technique

Algorithm :SSM- OPTICS(DB, eps, MinPts)

Input: : D = {T_1, T_2, …….Tn} is the dataset consisting of web user transactions where each T_i = {URLi1, URLi2, …..URLim} //URLs visited by user I.

Eps is the radius

MinPts is the Minimum number of neighborhood points .

P [0, 1]

Output: Set of Clusters C={c1,c2,c3…cn}

Method:

Step 1: For each point P of DB

Step2: Compute $S(T_i, T_j)$ using SSM Measure (equation 5).

Generate the neighbors of p whose distance is with in the

epsilon value(Eps) , say N=getNeighbors(p, eps)

Step3: If N>=MinPts, then mark p as core object ,

else mark p as NOISE.

Step 4: Add P to the priority queue.

Step 5; repeat this process until end of the database has reached.

VI. DATA PREPROCESSING

In this work we used MSNBC dataset (www.msnbc.com) founded in 1996 as a joint venture between Microsoft and NBC [MSNBC website]. This is a famous online news website with has different news subjects. For example: breaking news, extensive sources, advanced technology, original journalism and expansive content. The msnbc.com internet information server(IIS) always creates a log file with sequential list of web pages

VII.EXPERIMENTAL RESULTS

A. Example 1: DBSCAN clustering technique:

Experiments on Standard Data

TABLE 5.1 INTRA AND INTER LUSTER DISTANCE USING DBSCAN ON STANDARD DATA

DBSCAN	CLUSTERING RESULTS USING EUCLIDEAN				
No of samples	5000	10000	20000	30000	40000
No of clusters formed	83	123	156	115	191
Average inter cluster	4.7	4.906	5.123	7.213	
Average Intra cluster	4.26	4.021	4.0992	6.862	
	CLUSTERING RESULTS USING JACCARD				
No of samples	5000	10000	20000	30000	40000
No of clusters formed	99	114	147	135	197
Average inter cluster	4.281	4.317	5.213	6.153	6.297
Average Intra cluster	4.013	4.291	5.222	5.293	6.094
	CLUSTERING RESULTS USING COSINE				
No of samples	5000	10000	20000	30000	40000
No of clusters formed	96	123	156	115	191
Average inter cluster	4.6	6.367	7.214	8.135	7.218
Average Intra cluster	4.25	4.285	6.279	7.284	6.215
	CLUSTERING RESULTS USING FUZZY				

No of samples	5000	10000	20000	30000	40000
No of clusters formed	96	127	153	129	193
Average inter cluster	4.93	5.298	5.297	6.297	
Average Intra cluster	4.26	5.178	4.453	5.286	

5.2 Experiments On Synthetic Data

The results listed below considered arbitrarily with 200 records of web transactions from MSNBC.COM website. Table 5.1 shows the number of clusters formed using the DBSCAN clustering technique, and the average intra and inter cluster distance for the standard dataset MSNBC using various distance measures like Euclidean,Jaccard,Cosine,and Fuzzy similarity . Table 5.2,5.3,5.4,5.5 shows the inter cluster distance formed using Euclidean, jaccard, cosine and Fuzzy similarity measures... Table 5.6 shows the inter and intra cluster distance for the standard dataset, Table 5.7,5.8 shows the number of clusters formed using the OPTICS cluster technique and inter cluster distance is calculated using synthetic data . Table 5. 9 shows the average intra and inter cluster distance for the standard dataset .

TABLE 5.2. INTER CLUSTER DISTANCE FOR CLUSTERS FORMED USING DBSCAN ON MSNBC DATASET.

	C1	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*	*C10*	*C11*	*C12*	*C13*	*C14*	*C15*	*C16*	*C17*	*C18*	*C19*
C1	0	0.15	0.16	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.19	0.21	0.22	0.23	0.24	0.25	0.25	0.26
C2	0.15	0	0.13	0.13	0.14	0.13	0.13	0.14	0.15	0.16	0.17	0.18	0.21	0.22	0.23	0.24	0.25	0.26	0.27
C3	0.16	0.13	0	0.12	0.14	0.15	0.15	0.15	0.16	0.17	0.19	0.21	0.26	0.27	0.27	0.31	0.31	0.29	0.29
C4	0.16	0.13	0.12	0	0.18	0.18	0.18	0.19	0.19	0.19	0.21	0.21	0.22	0.23	0.24	0.24	0.25	0.26	0.26
C5	0.16	0.14	0.14	0.18	0	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.21	0.2	0.2	0.21	0.22	0.23
C6	0.16	0.13	0.15	0.18	0.16	0	0.16	0.17	0.18	0.21	0.22	0.23	0.24	0.24	0.25	0.25	0.25	0.25	0.26
C7	0.17	0.13	0.15	0.18	0.16	0.16	0	0.21	0.21	0.22	0.23	0.24	0.25	0.25	0.25	0.26	0.26	0.27	0.27
C8	0.17	0.14	0.15	0.19	0.16	0.17	0.21	0	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.21	0.21	0.22	0.25
C9	0.18	0.15	0.16	0.19	0.17	0.18	0.21	0.18	0	0.21	0.21	0.22	0.24	0.25	0.26	0.27	0.21	0.22	0.25
C10	0.18	0.16	0.17	0.19	0.17	0.21	0.22	0.18	0.21	0	0.22	0.23	0.23	0.24	0.24	0.25	0.25	0.27	0.27
C11	0.19	0.17	0.19	0.21	0.18	0.22	0.23	0.18	0.21	0.22	0	0.18	0.19	0.19	0.21	0.21	0.22	0.22	0.21
C12	0.19	0.18	0.21	0.21	0.18	0.23	0.24	0.18	0.22	0.23	0.18	0	0.19	0.14	0.14	0.21	0.24	0.24	0.25
C13	0.21	0.21	0.26	0.22	0.19	0.24	0.25	0.19	0.24	0.23	0.19	0.19	0	0.21	0.22	0.23	0.24	0.25	0.26

C14	0.22	0.22	0.27	0.23	0.21	0.24	0.25	0.19	0.25	0.24	0.19	0.14	0.21	0	0.25	0.26	0.26	0.27	0.27
C15	0.23	0.23	0.27	0.24	0.2	0.25	0.25	0.19	0.26	0.24	0.21	0.14	0.22	0.25	0	0.24	0.24	0.24	0.24
C16	0.24	0.24	0.31	0.24	0.2	0.25	0.26	0.21	0.27	0.25	0.21	0.21	0.23	0.26	0.24	0	0.19	0.23	0.22
0.17	0.25	0.25	0.31	0.25	0.21	0.25	0.26	0.21	0.21	0.25	0.22	0.24	0.24	0.26	0.24	0.19	0	0.19	0.2
C18	0.25	0.26	0.29	0.26	0.22	0.25	0.27	0.22	0.22	0.27	0.22	0.24	0.25	0.27	0.24	0.23	0.19	0	0.2
C19	0.26	0.27	0.29	0.26	0.23	0.26	0.27	0.25	0.25	0.27	0.21	0.25	0.26	0.27	0.24	0.22	0.2	0.2	0

TABLE 5.3: THE INTER CLUSTER DISTANCE USING JACCARD DISTANCE MEASURE.

cosine	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
C1	0	0.15	0.16	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.19	0.21	0.22	0.23	0.24	0.25	0.25
C2	0.15	0	0.13	0.13	0.14	0.13	0.13	0.14	0.15	0.16	0.17	0.18	0.21	0.22	0.23	0.24	0.25	0.26
C3	0.16	0.13	0	0.12	0.14	0.15	0.15	0.15	0.16	0.17	0.19	0.21	0.26	0.27	0.27	0.31	0.31	0.29
C4	0.16	0.13	0.12	0	0.18	0.18	0.18	0.19	0.19	0.19	0.21	0.21	0.22	0.23	0.24	0.24	0.25	0.26
C5	0.16	0.14	0.14	0.18	0	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.21	0.2	0.2	0.21	0.22
C6	0.16	0.13	0.15	0.18	0.16	0	0.16	0.17	0.18	0.21	0.22	0.23	0.24	0.24	0.25	0.25	0.25	0.25
C7	0.17	0.13	0.15	0.18	0.16	0.16	0	0.21	0.21	0.22	0.23	0.24	0.25	0.25	0.25	0.26	0.26	0.27
C8	0.17	0.14	0.15	0.19	0.16	0.17	0.21	0	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.21	0.21	0.22
C9	0.18	0.15	0.16	0.19	0.17	0.18	0.21	0.18	0	0.21	0.21	0.22	0.24	0.25	0.26	0.27	0.21	0.22
C10	0.18	0.16	0.17	0.19	0.17	0.21	0.22	0.18	0.21	0	0.22	0.23	0.23	0.24	0.24	0.25	0.25	0.27
C11	0.19	0.17	0.19	0.21	0.18	0.22	0.23	0.18	0.21	0.22	0	0.18	0.19	0.19	0.21	0.21	0.22	0.22
C12	0.19	0.18	0.21	0.21	0.18	0.23	0.24	0.18	0.22	0.23	0.18	0	0.19	0.14	0.14	0.21	0.24	0.24
C13	0.21	0.21	0.26	0.22	0.19	0.24	0.25	0.19	0.24	0.23	0.19	0.19	0	0.21	0.22	0.23	0.24	0.25
C14	0.22	0.22	0.27	0.23	0.21	0.24	0.25	0.19	0.25	0.24	0.19	0.14	0.21	0	0.25	0.26	0.26	0.27
C15	0.23	0.23	0.27	0.24	0.2	0.25	0.25	0.19	0.26	0.24	0.21	0.14	0.22	0.25	0	0.24	0.24	0.24
C16	0.24	0.24	0.31	0.24	0.2	0.25	0.26	0.21	0.27	0.25	0.21	0.21	0.23	0.26	0.24	0	0.19	0.23
C17	0.25	0.25	0.31	0.25	0.21	0.25	0.26	0.21	0.21	0.25	0.22	0.24	0.24	0.26	0.24	0.19	0	0.19
C18	0.25	0.26	0.29	0.26	0.22	0.25	0.27	0.22	0.22	0.27	0.22	0.24	0.25	0.27	0.24	0.23	0.19	0

TABLE 5.4: INTER CLUSTER DISTANCE USING COSINE SIMILARITY MEASURE

Cosine	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
C1	0	0.16	0.17	0.17	0.17	0.18	0.19	0.21	0.19	0.19	0.19	0.19	0.2	0.21	0.22	0.21	0.22	0.23
C2	0.16	0	0.17	0.18	0.17	0.17	0.18	0.18	0.17	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.19
C3	0.17	0.17	0	0.11	0.12	0.13	0.14	0.14	0.14	0.13	0.14	0.13	0.13	0.14	0.15	0.16	0.17	0.18
C4	0.17	0.18	0.11	0	0.16	0.13	0.19	0.17	0.17	0.17	0.16	0.17	0.18	0.17	0.18	0.17	0.17	0.19
C5	0.17	0.17	0.12	0.16	0	0.13	0.2	0.18	0.18	0.18	0.14	0.18	0.19	0.16	0.17	0.18	0.18	0.19
C6	0.18	0.17	0.13	0.13	0.13	0	0.21	0.18	0.18	0.18	0.16	0.17	0.16	0.15	0.16	0.18	0.16	0.17
C7	0.19	0.18	0.14	0.19	0.2	0.21	0	0.18	0.22	0.19	0.15	0.18	0.17	0.16	0.14	0.17	0.21	0.22
C8	0.21	0.18	0.14	0.17	0.18	0.18	0.18	0	0.23	0.2	0.16	0.17	0.18	0.18	0.17	0.17	0.18	0.18
C9	0.19	0.17	0.14	0.17	0.18	0.18	0.22	0.23	0	0.21	0.21	0.17	0.18	0.16	0.21	0.18	0.16	0.18
C10	0.19	0.16	0.13	0.17	0.18	0.18	0.19	0.2	0.21	0	0.22	0.19	0.19	0.17	0.22	0.18	0.18	0.19
C11	0.19	0.16	0.14	0.16	0.14	0.16	0.15	0.16	0.21	0.22	0	0.27	0.23	0.23	0.22	0.24	0.22	0.2
C12	0.19	0.16	0.13	0.17	0.18	0.17	0.18	0.17	0.17	0.19	0.27	0	0.18	0.17	0.18	0.17	0.17	0.19
C13	0.2	0.17	0.13	0.18	0.19	0.16	0.17	0.18	0.18	0.19	0.23	0.18	0	0.16	0.17	0.18	0.18	0.19
C14	0.21	0.17	0.14	0.17	0.16	0.15	0.16	0.18	0.16	0.17	0.23	0.17	0.16	0	0.16	0.18	0.16	0.17
C15	0.22	0.18	0.15	0.18	0.17	0.16	0.14	0.17	0.21	0.22	0.22	0.18	0.17	0.16	0	0.17	0.21	0.22
C16	0.21	0.18	0.16	0.17	0.18	0.18	0.17	0.17	0.18	0.18	0.24	0.17	0.18	0.18	0.17	0	0.18	0.18
C17	0.22	0.19	0.17	0.17	0.18	0.16	0.21	0.18	0.16	0.18	0.22	0.17	0.18	0.16	0.21	0.18	0	0.18
C18	0.23	0.19	0.18	0.19	0.19	0.17	0.22	0.18	0.18	0.19	0.2	0.19	0.19	0.17	0.22	0.18	0.18	0

TABLE 5.5 THE INTER CLUSTER TABLE USING SIMILARITY MEASURE FUZZY SIMILARITY MEASURE

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
C1	0	0.15	0.16	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19
C2	0.15	0	0.13	0.13	0.14	0.13	0.13	0.14	0.15	0.16	0.17
C3	0.16	0.13	0	0.12	0.14	0.15	0.15	0.15	0.16	0.17	0.19
C4	0.16	0.13	0.12	0	0.18	0.18	0.18	0.19	0.19	0.19	0.21

C5	0.16	0.14	0.14	0.18	0	0.16	0.16	0.16	0.17	0.17	0.18
C6	0.16	0.13	0.15	0.18	0.16	0	0.16	0.17	0.18	0.21	0.22
C7	0.17	0.13	0.15	0.18	0.16	0.16	0	0.21	0.21	0.22	0.23
C8	0.17	0.14	0.15	0.19	0.16	0.17	0.21	0	0.18	0.18	0.18
C9	0.18	0.15	0.16	0.19	0.17	0.18	0.21	0.18	0	0.21	0.21
C10	0.18	0.16	0.17	0.19	0.17	0.21	0.22	0.18	0.21	0	0.22
C11	0.19	0.17	0.19	0.21	0.18	0.22	0.23	0.18	0.21	0.22	0

TABLE 5.6 INTRA AND INTER LUSTER DISTANCE USING OPTICS ON STANDARD DATA

OPTICS	**CLUSTERING RESULTS USING EUCLIDEAN**				
No of samples	5000	10000	20000	30000	40000
No of clusters formed	94	126	149	141	187
Average inter cluster	5.275	6.271	7.312	8.156	6.216
Average Intra cluster distance	4.267	5.387	6.415	6.298	5.978
	CLUSTERING RESULTS USING JACCARD				
No of samples	5000	10000	20000	30000	40000
No of clusters formed	101	121	145	139	198
Average inter cluster	4.278	5.275	5.287	8.267	7.638
Average intra cluster	4.211	5.137	5.167	7.187	6.287
	CLUSTERING RESULTS USING COSINE				
No of samples	5000	10000	20000	30000	40000
No of clusters formed	99	127	148	127	187
Average Inter cluster	4.216	6.127	3.158	6.215	7.196
Average Intra cluster	3.156	4.178	3.170	4.218	5.218
	CLUSTERING RESULTS USING FUZZY				
No of samples	5000	10000	20000	30000	40000
No of clusters formed	87	126	139	143	186
Average inter cluster	3.217	4.128	5.126	7.219	6.218
Average Intra cluster	3.196	4.217	5.067	4.218	6.176

TABLE 5.7. INTER CLUSTER DISTANCE FORMED USING EUCLIDEAN DISTANCE MEASURE USING OPTICS FOR MSNBC DATASET

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
C1	0	0.16	0.17	0.17	0.17	0.18	0.19	0.21	0.19	0.19	0.19	0.19	0.2	0.21	0.22
C2	0.16	0	0.17	0.18	0.17	0.17	0.18	0.18	0.17	0.16	0.16	0.16	0.17	0.17	0.18
C3	0.17	0.17	0	0.11	0.12	0.13	0.14	0.14	0.14	0.13	0.14	0.13	0.13	0.14	0.15
C4	0.17	0.18	0.11	0	0.11	0.12	0.14	0.16	0.18	0.12	0.14	0.15	0.15	0.15	0.16
C5	0.17	0.17	0.12	0.11	0	0.17	0.16	0.17	0.17	0.16	0.16	0.16	0.17	0.17	0.18
C6	0.18	0.17	0.13	0.12	0.17	0	0.11	0.12	0.12	0.13	0.14	0.13	0.13	0.14	0.15
C7	0.19	0.18	0.14	0.14	0.16	0.11	0	0.19	0.18	0.19	0.17	0.19	0.2	0.21	0.22
C8	0.21	0.18	0.14	0.16	0.17	0.12	0.19	0	0.17	0.16	0.21	0.17	0.18	0.18	0.18
C9	0.19	0.17	0.14	0.18	0.17	0.12	0.18	0.17	0	0.21	0.22	0.17	0.18	0.18	0.22
C10	0.19	0.16	0.13	0.12	0.16	0.13	0.19	0.16	0.21	0	0.17	0.17	0.18	0.18	0.19
C11	0.19	0.16	0.14	0.14	0.16	0.14	0.17	0.21	0.22	0.17	0	0.16	0.14	0.16	0.15
C12	0.19	0.16	0.13	0.15	0.16	0.13	0.19	0.17	0.17	0.17	0.16	0	0.18	0.17	0.18
C13	0.2	0.17	0.13	0.15	0.17	0.13	0.2	0.18	0.18	0.18	0.14	0.18	0	0.16	0.17
C14	0.21	0.17	0.14	0.15	0.17	0.14	0.21	0.18	0.18	0.18	0.16	0.17	0.16	0	0.16
C15	0.22	0.18	0.15	0.16	0.18	0.15	0.22	0.18	0.22	0.19	0.15	0.18	0.17	0.16	0

TABLE 5.8: CLUSTERS FORMED USING JACCARD SIMILARITY MEASURE

Jaccard	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
C1	0	0.15	0.16	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.19	0.21	0.22	0.23	0.24	0.25	0.25
C2	0.15	0	0.13	0.13	0.14	0.13	0.13	0.14	0.15	0.16	0.17	0.18	0.21	0.22	0.23	0.24	0.25	0.26
C3	0.16	0.13	0	0.12	0.14	0.15	0.15	0.15	0.16	0.17	0.19	0.21	0.26	0.27	0.27	0.31	0.31	0.29
C4	0.16	0.13	0.12	0	0.18	0.18	0.18	0.19	0.19	0.19	0.21	0.21	0.22	0.23	0.24	0.24	0.25	0.26
C5	0.16	0.14	0.14	0.18	0	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.21	0.2	0.2	0.21	0.22
C6	0.16	0.13	0.15	0.18	0.16	0	0.16	0.17	0.18	0.21	0.22	0.23	0.24	0.24	0.25	0.25	0.25	0.25
C7	0.17	0.13	0.15	0.18	0.16	0.16	0	0.21	0.21	0.22	0.23	0.24	0.25	0.25	0.25	0.26	0.26	0.27
C8	0.17	0.14	0.15	0.19	0.16	0.17	0.21	0	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.21	0.21	0.22
C9	0.18	0.15	0.16	0.19	0.17	0.18	0.21	0.18	0	0.21	0.21	0.22	0.24	0.25	0.26	0.27	0.21	0.22
C10	0.18	0.16	0.17	0.19	0.17	0.21	0.22	0.18	0.21	0	0.22	0.23	0.23	0.24	0.24	0.25	0.25	0.27
C11	0.1	0.1	0.1	0.2	0.1	0.2	0.2	0.1	0.2	0.2	0	0.1	0.1	0.1	0.2	0.2	0.2	0.2

	9	7	9	1	8	2	3	8	1	2		8	9	9	1	1	2	2
C12	0.19	0.18	0.21	0.21	0.18	0.23	0.24	0.18	0.22	0.23	0.18	0	0.19	0.14	0.14	0.21	0.24	0.24
C13	0.21	0.21	0.26	0.22	0.19	0.24	0.25	0.19	0.24	0.23	0.19	0.19	0	0.21	0.22	0.23	0.24	0.25
C14	0.22	0.22	0.27	0.23	0.21	0.24	0.25	0.19	0.25	0.24	0.19	0.14	0.21	0	0.25	0.26	0.26	0.27
C15	0.23	0.23	0.27	0.24	0.2	0.25	0.25	0.19	0.26	0.24	0.21	0.14	0.22	0.25	0	0.24	0.24	0.24
C16	0.24	0.24	0.31	0.24	0.2	0.25	0.26	0.21	0.27	0.25	0.21	0.21	0.23	0.26	0.24	0	0.19	0.23
C17	0.25	0.25	0.31	0.25	0.21	0.25	0.26	0.21	0.21	0.25	0.22	0.24	0.24	0.26	0.24	0.19	0	0.19
C18	0.25	0.26	0.29	0.26	0.22	0.25	0.27	0.22	0.22	0.27	0.22	0.24	0.25	0.27	0.24	0.23	0.19	0

TABLE 5.9 SUMMARIZE OF AVERAGE INTRA CLUSTER DISTANCE FOR SSM-DBSCAN AND SSM-OPTICS

No of clusters	SSM-DBSCAN		SSM-OPTICS	
C1	4.33	0.360833	4.17	0.362609
C2	4.31	0.359167	3.84	0.333913
C3	4.69	0.390833	3.35	0.291304
C4	4.31	0.359167	3.54	0.307826
C5	4.59	0.3825	3.7	0.321739
C6	4.5	0.375	3.35	0.291304
C7	4.57	0.380833	4.04	0.351304
C8	4.38	0.365	3.83	0.333043
C9	4.38	0.365	4.15	0.36087
C10	4.64	0.386667	3.87	0.336522
C11	4.4	0.366667	3.56	0.309565
C12	4.29	0.3575	3.85	0.334783
C13	4.46	0.371667	3.74	0.325217
C14	4.73	0.394167	3.73	0.324348
C15	4.77	0.3975	4.04	0.351304
C16	4.96	0.413333	4.04	0.351304
C17	4.8	0.4	4.08	0.354783
C18	5.12	0.426667	4.19	0.364348
C19	5.27	0.439167	4.29	0.373043

C20	5.28	0.44	4.47	0.388696
C21	5.33	0.444167	4.35	0.378261
C22	5.44	0.453333	4.34	0.377391
C23	5.37	0.4475		
	4.33	0.360833	4.17	0.362609

VII. CONCLUSIONS

Web usage clustering is an important task in web mining in order to group similar sessions and identify web user access behavior. We have analyzed various density based clustering algorithms and compared their complexity and input parameters. Here in this paper , a novel similarity measure for sequential data is introduced and we proposed a enhanced density based clustering technique in purpose of finding the meaningful clusters in different databases.In our experiments, we compared the clustering characteristics of DBSCAN, OPTICS algorithm with the newly developed SSM-DBSCAN algorithm and SSM-OPTICS on the web sessions. We considered various similarity measures like Euclidean, Jaccard Coefficient, Fuzzy dissimilarity, and Cosine. In our work we considered both content and the order of content . We generated the clusters .The average of inter cluster and intra cluster using Levensthein distance are calculated. The results of DBSCAN are compared with results of SSM-DBSCAN and similarly , the results of OPTICS are compared with the results of SSM-OPTICS. The results of new similarity measure are compared with previously existing similarity measures like Euclidean , Jaccard, projected Euclidean, Cosine and fuzzy similarity .The SSM-DBSCAN average inter cluster distance , and Intra cluster distance is more when compared with the general DBSCAN , and the average inter cluster distance is more or less equal in SSM-OPTICS and OPTICS algorithm. Similarly the Intra cluster distance is minimum in in SSM-OPTICS then OPTICS. Finally showed behavior of clusters that made by enhanced DBSCAN clustering technique and OPTICS algorithm on a sequential data in a web usage domain, and takes less time requirement by the way of explanations and list of conclusions. This experiment shows the efficiency of the new algorithm.

REFERENCES

[1] Aoying.Z, Shuigeng.Z, “Approaches for scaling DBSCAN algorithm to large spatial database”, Journal of Computer Science and Technology , Vol. 15(6), 2000, pp. 509–526).

[2] Aggarwal.C, Han.J, Wang.J, Yu.P.S,”A Framework for Clustering Evolving Data Streams”, Proc. Of Int. Conf. on Very Large DataBases, Berlin, Germany, Vol.32,Sept. 2003,pp.(5-14).

[3] Cooley.R , Mobasher B.,Srivastava, Data preparation for mining world wide web browsing patterns, Knowledge and Information Systems, Vol 1, 1999, pp.(5-32).

[4] Cao.F, Estery .M, Qian .W ,Density based clustering over an eveloving data stream with noise, SDM SIAM conference on Data Mining, **Vol.**32(2),pp. **pp**.(5--14.),2006

[5] Chen.Y ,Density based clustering for real time stream data, vol pp.International conference on Knowledge discovery and databases Vol. 26, 2007.pp.(323-369),

[6] Ester.M, Kriegel.H.P, Sander.J, and Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1996, pp. (226-231).

[7] Ester.M., Kriegel.H.P., Sander.J, Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, 1996, pp. (226–231).

[8] Forster.A, Murphy A.L, CLIQUE: Role-Free Clustering with Q-Learning for Wireless Sensor Networks, Distributed Computing Systems, 29th IEEE International Conference, 2009,

[9] Guha.S., Rastogi.R., Shim.K, an efficient clustering algorithms for large databases, In Proceeding ACM SIGMOD International Conference on Management of Data, Seattle, WA, 1998, pp. (73–84).

[10] Guha.s, Mishra.n, Motwani.r, Callaghan.l, Clustering data streams. In Proceedings of Computer Science. IEEE, Vol.16(10),November 2000,pp.(1391—1399)..

[11] Guha.S, Meyerson.A,. Mishra.N, Motwani.R, and Callaghan.L, Clustering Data Streams: Theory and Practice TKDE special issue on clustering, Vol.15, 2003.

[12] Goldman.S, Zhou.YEnhancing supervised learning with unlabeled data,*Proc. of ICML*, 2000, pp.(92–100).

[13] Ganti.V, Johannes Gehrke, Raghu Ramakrishnan, Mining Data Streams under Block Evolution, SIGKDD Explorations Vol.3(2), 2002.

[14] Gaber, Zaslavsky.M, Krishnaswamy.A, Towards an Adaptive Approach for Mining Data Streams in Resource Constrained Environments, the Proceedings of Sixth International Conference on Data Warehousing and Knowledge Discovery – Industry Track , August – 3 September, Lecture Notes in Computer Science (LNCS),Springer,Verlag, 2004.

[15] Mobasher.B, Colley.R, Srivastava.J, Automatic personalization based on web usage mining, Commun. ACM 43, 8 ,2000, pp.(142-151).

AUTHORS PROFILE

1. Ms K.Santhisree,Associate Professor,Dept of computer science and Engg, Jawaharlal Nehru Technological university, Hyderabad. She is pursuing her Ph.D in the area of Data Mning. Her research interests include Data Mniing,Data strucutres, Design and analysis of Algorithms, Information Retrieval systems, and she is having 10 years of Teaching experience. She is an coordinator to school of Continuous and distance education.

2. Dr A.Damodaram, Professor of Department of Computer science nd Engineering at JNTU Hyderabad.. He is presently a Director of School of continuous and dsitance education. He was a Vice –principal to JNTUH College of Engg, JNTU Hyderabad. His research interests include Networks,computer ommunications, Information security systems.He has 19 years of Tteaching experience.