Case Study : Comparative Analysis: On Clustering of Sequential Data Streams USING Optics and ICA

K. Santhi Sree, PhD
Jawaharlal Nehru Technological University
Hyderabad

R. Kranthi Kumar
Vensi Streamlined Solutions
Hyderabad

ABSTRACT
Clustering on web usage data is useful to identify what users are exactly looking for on the world wide web, lik e user traversals, users behavior and their characteristics, which helps for Web personalization. Clustering web sessions is to group them based on similarity and consists of minimizing the Intra-cluster similarity and maximizing the Inter-group similarity. In the past there exist multiple similarity measures like Euclidean, Jaccard ,Cosine ,Manhanttan, Minkowski, and many to measure similarity between web patterns. In this paper, we enhanced Icremental clustering algorithm (ICA) based on OPTICS. Experiments are performed on MSNBC.COM website (free online news channel), on sequential data streams in the context of clustering in the domain of Web usage mining. Specially, we present a detailed comparison of ICA and OPTICS and the results illustrate that ICA is much more suitable for clustering the dynamic datasets. The Inter-cluster and Intra-cluster distances are computed using Average Levensthien distance (ALD) to demonstrate the usefulness of the proposed approach in the context of web usage mining. This new enhanced (ICA algorithm)has good results when compared with existing OPTICS clustering technique , and provided good time requirements of the newly developed algorithms.

Keywords
Sequence Mining, Clustering ,Density Based Clustering(optics). Data Mining, Clustering, similarity measures, Web Personalization,

1. INTRODUCTION
1.1 Clustering
Clustering is a process of categorizing the data into multiple clusters where all the patterns lying in one cluster are similar to one another and dissimilar when compared to the patterns lying in the other cluster . Different types of clustering techniques are partitioning ,Hierarchical, Density-based ,Grid-based and Model–Based algorithms. Types of Density based clustering techniques are DbScan, Optics and Denclue. Here in this work, we are concentrating on OPTICS clustering technique.

1.2 Similarity Measures
Similarity measure are used to find out how similar are two sequences are. In the history many similarity measures exist, and they are Euclidean, Jaccard, Cosine, Manhanttan and Minkowski measures. These similarity measures are either vector based or frequency based. The Euclidean distance between sequences $S_1=(p_1, p_2,..., p_n)$ and $S_2=(q_1, q_2,..., q_n)$ is defined as

$$Sim(S_1, S_2) = \sqrt{(S_{1_1} - S_{2_1})^2 + (S_{1_2} - S_{2_2})^2 + \cdots + (S_{1_n} - S_{2_n})^2} = \sqrt{\sum_{i=1}^{n} (S_{1_i} - S_{2_i})^2} \quad \text{----(1)}$$

Jaccard similarity measure is defined as the ratio of the intersection of items between the two sequences to the union of items of the two sequences.

$$(Sim(S_1, S_2)) = \frac{S_1 S_2}{|S_1|^2 + |S_2|^2 - S_1 S_2} \quad \text{---- (2)}$$

Cosine similarity measure is the angle between two vectors. The cosine measure is given by

$$Sim(S_1, S_2) = \frac{\sum_{i=1}^{n} (S_1 \times S_2)}{\sqrt{\sum_{i=1}^{n} (S_{1i})^2} \times \sqrt{\sum_{i=1}^{n} (S_{2i})^2}} \quad \text{---------(3)}$$

2. EXISTING METHODOLOGY
In the existing work the sequences are converted to intermediate representations and the similarity between any two sequences is calculated using any of the similarity measures like Euclidean. OPTICS clustering technique can be applied for clustering. While computing similarity between sequences they either consider the content /information or the order information .

Algorithm : OPTICS(DB, Eps, MinPts)
Input: A database D with N samples.
{Dataset D with N objects, epsilon(eps) the radius, and minpts, i.e.. the number of minimum points and C the cluster}
Output: set of Clusters C={c1,c2,c3…cn}
Method:
Step 1: For each point P of DB
Step 2 : N = regionQuery(P, eps) .
Step 3: If sizeof(N) < MinPts , mark P as NOISE.

Step 4: If N>=MinPts, then mark p as core object

Step 5: Add P to the priority queue.
Step 6; Repeat steps 1,2,3 ,4 until end of the database DB has reached

Fig 1. Algorithm for Optics Clustering

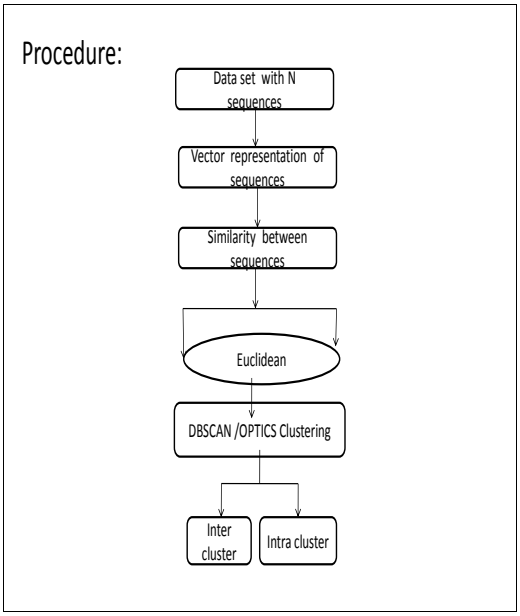

Fig 2. Exiting Work Procedure

Example:
Step1: Consider a set of 10 sample sequences (Transactions) randomly from the MSNBC dataset. In the current work, the sequences has to be converted to vector representations. The entire set contains multiple categories of news like {on-air, misc, news, sports bbs, frontpage, local, weather, travel, opinion, msn-news, business etc}. In each sequence presence of category of news is taken as 1 and absence as 0.The vector representation of the sequences is In Table 1 and 2 the rows indicate the transactions {T1,T2,T3,T4,T5,T6,T7,T8,T9,T10} and the columns indicate the category id .i.e news id. In the first sequence for example , on-air is present, misc is present so the particular category id represented as 1 and remaining categories are taken as 0.

Table 1. Vector Representation Of Sequences

Transaction × category id	1	2	3	4	5	6	7	8	9	10	11	12
T1	1	1	0	0	0	0	0	0	0	0	0	0
T2	0	0	1	1	0	0	1	0	1	0	0	0
T3	0	0	0	0	1	0	0	0	0	0	0	0
T4	0	0	1	1	0	1	1	0	0	0	0	0
T5	1	0	0	1	0	0	0	1	0	0	0	0
T6	1	0	0	0	1	0	0	0	1	0	0	0
T7	0	0	1	0	1	1	0	0	0	0	0	0
T8	0	0	0	0 -	1	1	0	0	0	0	0	0
T9	0	0	1	0	0	0	0	0	1	1	1	0
T10	0	0	1	0	1	1	0	0	0	0	0	1

In Table 5.2, For example ,consider the first sequence/Transaction, on-air is present, whose frequency is 2, misc is present and its frequency is 4.so the particular category id 1 and 2 are represented with its frequency 2 and 4 respectively and remaining category id's are considered as 0.

Table 2. Frequency Representation Of Sequences

Transaction × Category id	1	2	3	4	5	6	7	8	9	10	11	12
T1	2	4	0	0	0	0	0	0	0	0	0	0
T2	0	0	0	0	3	1	0	0	0	0	0	0
T3	0	0	0	0	0	6	0	0	0	0	0	0
T4	0	0	2	1	0	2	1	0	0	0	0	0
T5	1	0	0	1	0	0	0	3	0	0	0	0
T6	4	0	0	0	1	0	0	0	1	0	0	0
T7	0	0	1	0	2	3	0	0	0	0	0	0
T8	0	0	0	0	1	5	0	0	0	0	0	0
T9	0	0	2	0	0	0	1	2	1	0	0	0
T10	0	0	2	0	1	2	0	0	0	0	1	0

Table 3. sequence similarity matrix using Euclidean Measure

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	—	2.44	2.23	2.44	1.7321	1.4142	2.23	2	2.44	2.23
T2	2.44	—	2.23	1.4142	2.23	2.23	2.23	2.23	2	2.44
T3	2.23	2.23	—	2.23	2	1.4142	1.4142	1	2.23	1.73
T4	2.44	1.4142	2.23	—	2.23	2.64	2	2	2.44	2
T5	1.732	2.23	2	2.23	—	2	2.44	2	2.64	2.64
T6	1.414	2.23	1.414	2.64	2	—	2	1.7321	2.44	2.44
T7	2.23	2.23	1.414	2	2.44	2	—	1	2.44	1
T8	2	2.23	1	2	2	1.7321	1	—	2.44	1.41
T9	2.44	2	2.23	2.44	2.64	2.44	2.44	2.44	—	2.44
T10	2.23	2.44	1.73	2	2.64	2.44	1	1.41	2.44	—

Step3:

In Table 3 indicates a N×N Similarity matrix is calculated where rows and columns indicate the Transactions{ T1,T2,T3,T4,T5,T6,T7,T8,T9,T10}.For example similarity(T1,T2)=2.44 .i e similarity between the two sequences T1,T2 is 2.44. If the two sequences say T1,T2

are similar , the similarity(T1,T2)=0. If they are more dissimilar , the similarity ratio increases. For example, the similarity between the sequences (T1,T5)=1.732, which means the two sequences seem to be more similar. The similarity between the sequences (T5,T9)=2.64, which shows the two sequences seems to be more dissimilar.

Step 4: Applying OPTICS clustering algorithm:

Clusters formed are

- C1={T3,T5,T6,T8}
- C3={T1,T5,T6,T7,T8,T9}
- C5={T1,T3,T6}
- C6={T1,T3,T5,T7,T8}
- C7={T3,T6,T8,T10}
- C9={NOISE}
- C2={NOISE}
- C8={T1,T3,T4,T6,T7} and
- C10={T3,T7,T8}

Applying ICA clustering algorithm, the clusters formed are

- C1={ T1,T3,T5,T6,T8,T7,10}
- C2={T2}
- C3={T4}

3. PROPOSED WORK PROCEDURE

Enhanced Incremental Clustering algorithm: The work concentrates on Clustering techniques on data streams in the domain of web usage data . Euclidean similarity measure is used to measure similarity/distance between two sequences and experiments are conducted on various clustering techniques using ,Optics, and ICA . In all the experiments the running time of the new algorthim (ICA) is best compared to the earlier similarity measures. Figure 3 shows the proposed framework.

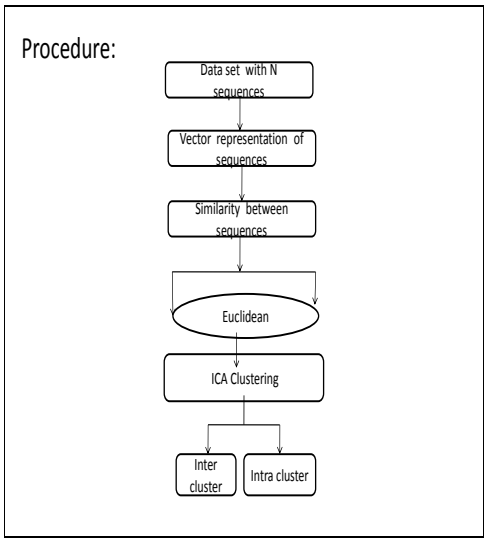

Fig 3: Proposed Work Procedure

4. EXPERIMENTAL RESULTS

4.1 Web Navigation Dataset for Testing

MSNBC is a famous online news website with has different news subjects. There are 17 categories of news likefrontpage,news,tech,local,opinion,onair,weather,health,living,business,sports,summary,bbs,travelmisc,msn-news, and msn-sports. Web Navigational dataset is considered in Table 4.

Table 4. Web Navigational Dataset

Sequence	
T1	on-air misc misc misc on-air misc
T2	news sports tech local sports ,sports
T3	Sports bbs bbs bbs bbs bbs bbs
T4	frontpage frontpage sports news news local
T5	on-air weather weather weather sports,sports
T6	on-air on-air on-air on-air tech bbs
T7	frontpage bbs bbs frontpage frontpage news
T8	frontpage frontpage frontpage frontpage frontpage bbs
T9	news news travel opinion opinion msn-news
T10	frontpage business frontpage news news bbs

4.2 Optics and ICA Experiments on Standard web Navigational Dataset

Considered transactions of varying sizes of 5000, 10000,20,000,30000,40000 from MSNBC dataset. Table 5 shows the number of clusters formed by applying the existing Optics and proposed ICA . Using the similarity measure like Euclidean, Inter cluster similarity and Intra cluster similarity are calculated.

Table 5. Inter and Intra cluster distance for OPTICS and ICA

OPTICS -**Clustering Results Using Euclidean**					
No of Samples	5000	10000	20000	30000	40000
No of clusters formed	82	124	155	116	189
Inter cluster	4.5	4.9	5.124	6.893	6.989
Average inter cluster	0.056	0.037	0.031	0.061	0.039
Average Intra cluster	4.27	4.000	4.989	6.867	5.896
ICA- Clustering Results Using Euclidean					
No of samples	5000	10000	20000	30000	40000
No of clusters formed	96	123	156	115	191
Inter cluster	4.6	6.367	7.214	8.135	6.721
Average Inter cluster	0.047	0.051	0.039	0.070	0.035
Average Intra cluster	4.25	4.285	6.279	7.284	6.912

5. TIME REQUIREMENTS

Experiments were performed on the above mentioned dataset of varying sizes ,to see the performance of existing and proposed clustering algorithms. The number of clusters formed using by these for varying sizes of 5000, 10000, 20000, 30000 and 40000 transactions are recorded. The execution time taken for these varying sizes of samples are also recorded in table 6.

Table 6. Time Requirements Of OPTICS and ICA

OPTICS					
Size of sequences	5000	10000	20000	30000	40,000
No of clusters	94	126	149	141	187
Time taken in seconds	1566	1879	3643	3218	4982
ICA					
Size of sequences	5000	10000	20000	30000	40,000
No of clusters	96	127	153	129	131
Time taken in seconds	785	1638	2064	1679	1875

6. CONCLUSIONS

Considered arbitrarily web transactions from the MSNBC dataset and performed the experiments on Clustering algorithms.. We used previously existing /similarity measure namely Euclidean. For good clustering algorithm, the intra cluster distance should be minimum. We generated clusters using OPTICS and ICA .Then using OPTICS and ICA, clusters are generated .Comparing OPTICS and ICA ,the inter cluster similarity is maximum in ICA. For example in OPTICS for 5000 samples ,the time taken for execution are 1566,1879,3643,3218,4928 respectively. The time taken to execute the algorithm ICA is less when compare to other clustering techniques . A variety of experiments are performed in the context of clustering. on a sequential data in a web usage domain.. This experiment shows that in addition to the content if Sequential Information is also added it improves the quality /accuracy of the clustering. So Sequential information is important as well as Content information is also important.

6.1 Future Work

- We extend our work in future to other clustering techniques and to other domains as well..
- The time complexities of the proposed algorithms can be improved further.

7. REFERENCES

[1] Aggarwal.C, Han.J, Wang.J, Yu.P.S, "A Framework for Projected Clustering of High Dimensional Data Streams", Proc. 2004 Int. Conf. on Very Large Data Bases, Toronto, Canada, pp.(852-863), 2004.

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

[3] Chen Song-Yu, O'Grady2,O'Hare, Wei Wang, **"**A Clustering Algorithm Incorporating Density and Direction", IAWTAC ,IEEE 2008.Deepak P, Shourya Roy IBM India Research Lab, OPTICS on Text Data: Experiments and Test Results.

[4] Cooley.R,Mobasher. B,Srivastava.J, "Web mining: Information and pattern discovery on the world wide web", 9th IEEE Int. Conf. Tools AI .

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

[6] K.Santhisree, Dr A.Damodaram, 'SSM-DBSCAN and SSM-OPTICS : Incorporating a new similarity measure for Density based Clustering of Web usage data". International Journal on Computer Science and Engineering (IJCSE),Vol.3(9),PP.(3170-3184)September 2011,India.

[7] K.Santhisree," SSM-DENCLUE : Enhanced Approach for Clustering of Sequential data: Experiments and Test cases, International Journal of Computer Applications,Vol.96(6),pp.(7-14),June 2014. Published by Foundation of Computer Science, New York, USA.