OPTICS on Sequential Data: Experiments and Test Results

K.Santhisree
Dept. of Computer science
JNT University (JNTUH)
Hyderabad, India

Dr A.Damodaram
Prof& Director of SCDE
Dept. of Computer science
JNT University (JNTUH)

ABSTRACT
The Web has enormous, various and knowledgeable data for data mining research. Clustering web usage data is useful to discover interesting patterns pertaining to user traversals, behaviour and their usage characteristics. Moreover, users accesses web pages in an order in which they are interested and hence incorporating sequence nature of their usage is crucial for clustering web transactions. In this paper we present OPTICS ("Ordering Points To Identify the Clustering Structure") algorithm to find density based clusters on a web usage data on MSNBC.COM website which is a free news data website with so different categories of news).The clusters are generated by OPTICS algorithm . The average of inter cluster and intra cluster are Calculated. the results are compared with different similarity measures like Euclidean , Jaccard, projected Euclidean, cosine and fuzzy similarity Finally showed behavior of clusters that made by OPTICS algorithm on a sequential data in a web usage domain. we performed a variety of experiments in the context of density based clustering , quantify our results by the way of explanation s and list conclusions.

Keywords
Clustering algorithm OPTICS, Ordering Points To Identify the Clustering Structure, Sequence mining. Average Inter cluster, Intra cluster.

1. INTRODUCTION
The web is a huge database for research about relationship between objects, people, socials, companies, relations, marketing, management, knowledge and etc. Clustering is a one of the ways to collecting subsets of data that have some common attributes and find hidden patterns to create knowledge from databases. Different types of data clustering are: Hierarchical, Partitional, Density-based, Sub-space clustering and etc. In this paper we use a density-based clustering algorithm that name is OPTICS algorithm to clustering a dataset from msnbc.com website. Because this algorithm is a density based algorithm and our data is a density based data. First of all we downloaded data (around 40'000 records) from the msnbc.com website and then created a dataset file. We did some preprocessing on the dataset to extract impossible data combinations for example it is possible that some attributes have never been used by the users. After accredit from the data we apply OPTICS algorithm on the new dataset to cluster the data. Finally use Euclidean distance measure, projected Euclidean distance, cosine similarity and Fuzzy dissimilarity to compare the results for intra cluster and inter cluster analyze and visualize data and graphs.

2. RELATED WORK
There are some other papers which applied OPTICS clustering with noise on the different datasets and analyze on various ways. "M Ankerst, M. Breunig, H.Kriegel, J.Sander" [7] introduce OPTICS algorithm on density based clustering structure. In [5] they showed how to generate appropriate distance information about compressed data points, and how to adapt the graphical representation of the clustering result with OPTICS algorithm. In another experience [1] Deepak, Roy previews various densities based clustering algorithm and describe the performance measures and feature selection techniques. Ester & Kriegel & Xu (KDD-96) work has worked on large spatial database with noise in [6]. They using synthetic data and real data of the SEQUOIDA 2000 benchmark and discover clusters of arbitrary shape. In [2] Dimitris K. Tasoulis, Gordon Ross, and Niall M. Adams extend OPTICS algorithm to streaming data model and visualise clusters and finally demonstrate the behaviour of OPTICSTREAM. The discovery of clusters from database updates is a problem that Parthasarathy, Zaki, Ogihara, Dwarkadas [8] has worked on it. They presented a method for incremental and interactive frequent sequence mining with SPADE algorithm. On paper [4] "Masson" and "Denoeux" explained about fuzzy dissimilarity and showed where dissimilarities are expressed as intervals or fuzzy numbers. "Hanzhou" and "Zhejiang" [3] worked on a protein sequence data with OPTICS algorithm and create a new algorithm with name of SEQOPTICS.

3. OPTICS
The OPTICS algorithm (Ordering Points To Identify the Clustering Structure) algorithm designed by Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel and Jörg Sander. Its basic idea is similar to DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to specify clusters and noises of a special database. OPTICS defines a cluster with the base of density. 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"*). 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"*. 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"*th point 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. A noise has special attributes that is not common with other noises and clusters. This process should be continued for all points to specify whether a point is cluster, e-neighborhood or a noise. Then we can specify clusters and noises with OPTICS algorithm.

4. DATA PREPROCESSING
The msnbc.com founded in 1996 as a joint venture between Microsoft and NBC [MSNBC website]. It's one of the biggest

online news organizations. Breaking news, original journalism, extensive sources, advanced technology and expansive content can find on that website. The msnbc.com internet information server (IIS) can create a log file with sequential list of pages that each user saw on msnbc.com.

In this paper we change an open source application which is developed by C++ with the use of some free libraries for calculation and visualization. After getting inchmeal data from the MSNBC website we added it to a file and did some preprocessing on it. On preprocessing level we must extract attributes that have not been used by any user because such an attribute causes infinity conditions to happen in calculations which mean we cannot have a cluster that has never been used by any user. When removed never used attributes we make a new dataset with list of operator attributes selection. After extracting unused attributes from the dataset a new dataset with a list of properly selected attributes was created which is a text file with 40'000 records of users. From the new dataset a two dimensional matrix was created in the memory. For the first step an *"m×n"* matrix must be created where *"m"* is the number of attributes and the *"n"* is number of dataset records. Then subtract of the mean for each attributes and calculate covariance the matrix. Finally we must calculate the Eigen vectors and Eigen values of the covariance matrix. Finally OPTICS algorithm is used to clustering the dataset and creates hidden patterns.

5. EXPERIMENTAL EVALUATION

The results listed below are produced using a home computer with T8100 processor, 2GB random access memory and Windows Vista as the operating system. The clusters are generated varying the epsilon values from 0.1 to 0.9. Table1 is a inter cluster table generated using projected Euclidean with epsilon value 0.3. Table 2 represents the average inter cluster distance calculated using Euclidean, projected Euclidean, Jaccard, cosine and fuzzy. The mean standard deviation and error rates are calculated.

Table 1: The Inter cluster distance

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**	**C19**
C1	0.00	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	0.00	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	0.00	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	0.00	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	0.00	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.18	0.19	0.21	0.20	0.20	0.21	0.23
C6	0.16	0.13	0.15	0.18	0.16	0.00	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	0.00	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	0.00	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.19	0.21	0.22	0.25
C9	0.18	0.15	0.16	0.19	0.17	0.18	0.21	0.18	0.00	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	0.00	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	0.00	0.18	0.19	0.19	0.19	0.21	0.21	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	0.00	0.19	0.14	0.21	0.14	0.21	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	0.00	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	0.00	0.25	0.26	0.26	0.27	0.27
C15	0.23	0.23	0.27	0.24	0.20	0.25	0.25	0.19	0.26	0.24	0.21	0.14	0.22	0.25	0.00	0.24	0.24	0.24	0.24
C16	0.24	0.24	0.31	0.24	0.20	0.25	0.26	0.21	0.27	0.25	0.21	0.14	0.21	0.21	0.23	0.26	0.24	0.00	0.19	0.23	0.22
0.17	0.25	0.25	0.31	0.25	0.21	0.25	0.26	0.21	0.25	0.21	0.21	0.25	0.22	0.24	0.24	0.26	0.24	0.19	0.00	0.19	0.20
C18	0.25	0.26	0.29	0.26	0.22	0.25	0.27	0.22	0.22	0.22	0.27	0.22	0.24	0.25	0.27	0.24	0.23	0.19	0.19	0.00	0.20
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	0.24	0.22	0.20	0.20	0.20	0.00

Graph 1: The Inter cluster line graph

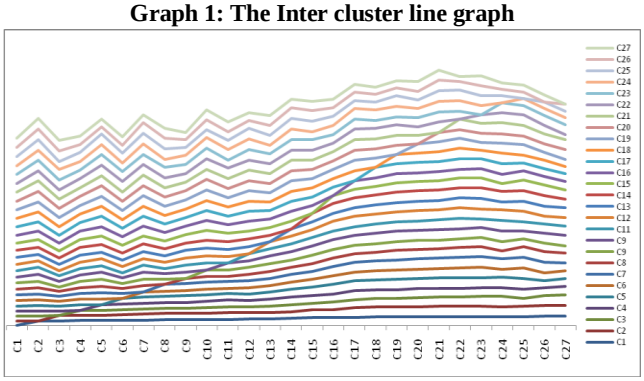


Table 2: The average of Inter cluster distance for different *"EPS"*

Eps	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AVG	0.126	0.128	0.160	0.150	0.150	0.143	0.158	0.162

Graph 2: The avarage of inter cluster bar graph

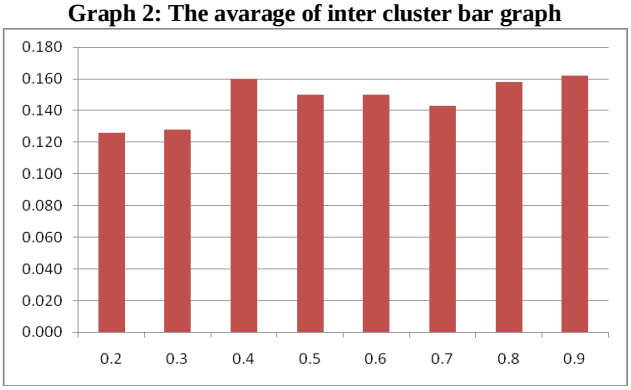


Table 3: Inter cluster distance

	mean	standard deviation	error
EUCLIDEAN	0.22	0.006022	
JACCARD	0.202	0.0072	
Projected Euclidean distance	0.142	0.0055	
Cosine similarity	0.16	0.0052	
Fuzzy dissimilarity	0.176	0.0061	
	0.16	0.0052	
	0.202	0.0072	
	0.144	0.0024	
	0.159	0.0036	

Graph 3: The Intra cluster line graph

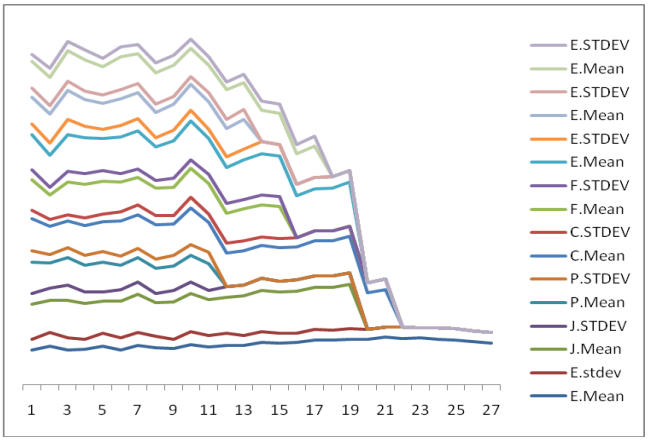
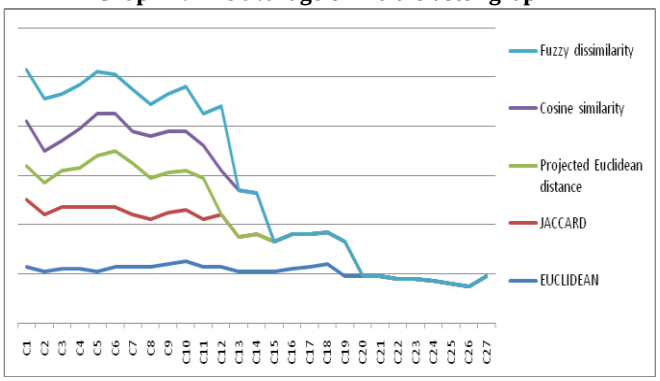


Table 4: The average of Intra cluster distance

	EUCLIDEAN	JACCARD	Projected Euclidean distance	Cosine similarity	Fuzzy dissimilarity
C1	0.23	0.27	0.14	0.18	0.21
C2	0.21	0.23	0.13	0.13	0.21
C3	0.22	0.25	0.15	0.12	0.19
C4	0.22	0.25	0.16	0.16	0.18
C5	0.21	0.26	0.21	0.17	0.17
C6	0.23	0.24	0.23	0.15	0.16
C7	0.23	0.21	0.21	0.13	0.17
C8	0.23	0.19	0.17	0.17	0.13
C9	0.24	0.21	0.16	0.17	0.15
C10	0.25	0.21	0.16	0.16	0.18
C11	0.23	0.19	0.17	0.13	0.13
C12	0.23	0.21	-	0.18	0.26
C13	0.21	0.14	-	0.19	-
C14	0.21	0.15	-	0.17	-
C15	0.21	0.12	-	-	-
C16	0.22	0.14	-	-	-
C17	0.23	0.13	-	-	-
C18	0.24	0.13	-	-	-
C19	0.19	0.14	-	-	-
C20	0.19	-	-	-	-
C21	0.19	-	-	-	-
C22	0.18	-	-	-	-
C23	0.18	-	-	-	-
C24	0.17	-	-	-	-
C25	0.16	-	-	-	-
C26	0.15	-	-	-	-

C27	0.19	-	-	-	-

Graph 4: The avarage of Intra cluster graph



6. CONCLUSION

Data mining and its sibling sequence mining are special processes which their goal is prediction. By studying clusters' behaviour it is possible to estimate the next value. In this research we use a dataset that describes the page visits of users who visited msnbc.com. we adopted a clustering OPTICS algorithm which is used to create clusters from data in the dataset then inter and intra cluster values are evaluated by using Euclidean distance measurement, projected Euclidean distance, cosine similarity and Fuzzy dissimilarity to find the similarity between cluster and the results are visualized graphically which helps in predicting the user behavior .

7. REFERENCES

[1]. "Deepak P, Shourya Roy" IBM India Research Lab, "OPTICS on Text Data: Experiments and Test Results".

[2]. "Dimitris K. Tasoulis, Gordon Ross, and Niall M. Adams "Department of Mathematics Imperial College London, "Visualising the Cluster Structure of Data Streams".

[3]. "Hanzhou, Zhejiang", "SEQOPTICS: A Protein Sequence Clustering Method", Computer and Computational Sciences, 2006. IMSCCS '06. First International Multi-Symposiums on.

[4]. "M. Masson, T. Denoeux", **"**Multidimensional scaling of fuzzy dissimilarity data", 2002, ISSN: 0165-0114.

[5]. "Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander", "Fast Hierarchical Clustering Based on Compressed Data and OPTICS" Proc. 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), Lyon, France.

[6]. Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proc. 2 International Conference on Knowledge Discovery and Data Mining (KDD-96). pp.226-231.

[7]. "Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander", "OPTICS: Ordering Points To Identify the Clustering Structure" Proc. ACM SIGMOD'99 Int. Conf. on Management of Data, Philadelphia PA, 1999.

[8]. Srinivasan Parthasarathy, Mohammed J. Zaki, Mitsunori Ogihara and Sandhya Dwarkadas, "Incremental and Interactive Sequence Mining". Proc. in 8th ACM International Conference Information and Knowledge Management. Nov 1999.

8. AUTHORS PROFILE

Ms K.Santhisree is presently working as Associate Professor, Department of Computer science, JNTU, Hyderabad. She has 10 years of teaching experience in the area of computer science. Her areas of Interest are Data mining, cloud computing , information Retrieval systems, Data structures and Design of algorithms.

Dr A. Damodaram is a professor in Department of computer science in Jawaharlal Nehru technological university Hyderabad. He is presently a director of SCDE. He has 20 years of teaching experience his interested areas are software engineering, computer networks, image processing.