International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 481-486

SEQDBSCAN : A NEW SEQUENCE DBSCAN ALGORITHM FOR CLUSTERING OF WEB USAGE DATA

K.Santhisree[1], D.Nagarjuna Devi[2] & K. Vijay Bhaskar[3]

Web is a vast area for data mining research. It is used in finding the user access patterns from web access log. User page visits are sequential in nature In this paper,I proposed a new clustering algorithm, SeqDBSCAN for clustering sequential data.. we adopted a similarity preserving function called sequence and set similarity measure S^3M that captures both the order of occurrence of page visits as well as the content of pages we conducted experiments comparing the results of SeqDBSCAN with other similarity measures S3M, Euclidean and Jaccards. The clusters resulting from these measures are computed using a cluster validation technique called Average levensthein distance(ALD).. Based on these results, We tested the new algorithm on dataset namely, MSNBC dataset and proved that the inter cluster similarity is high in S^3M when compared to the Euclidean and Jaccards distance measures and a set of experiments are conducted to investigate whether clustering performance is affected by different sequence representations, and different distance measures and other factors like number of web pages, similarity between clusters, number of user sessions , number of clusters to form.

Keywords: Sequence clustering ,web usage data, similarity measures,Inter-cluster similarity., Average Levensthein distance.

1. Introduction

Firstly we use a s^3m measure which was introduced earlier(pradeep kumar, Bapi, Krishna). secondly we compare the results of SeqDBSCAN algorithm with the other similarity measures Euclidean and Jaccards. Based on the results we design a new SeqDBSCAN algorithm. Finally we validate the clusters where ALD is used to calculate the Intra cluster and Inter cluster similarity. In the next session we discuss the related work on web personalization. Then we review the literature related to the set similarity and sequence similarity., and other similarity measures like Euclidean and Jaccards. And on clustering analysis. Then comes the discussion and preprocessing of msnbc dataset. then we present the clustering of web usage data using DBSCAN with Euclidean, Jaccards and S^3M.

2. Related Work

Web mining is the use of Data mining techniques to extract information from web documents . web usage mining is an active topic to the researchers for database management, Artificial intelligence and Information systems.

Many Data mining techniques such as Association rule mining, Sequence pattern mining, Cluster analysis are adopted to improve the scalability and usability of web mining techniques. Generally there are two methods of clustering techniques performed on the web usage data like data-user transaction and web page clustering (mobhasheer, 2000). (pradeep kumar, Krishna and Bapi 2007) developed SeqPAM for web personalization. (mosabheer, et al 2000) used developed automatic personalization of a website based on the web usage data. They clustered the web usage data based on cosine similarity measure. Web patterns were extracted from web logs using many data mining techniques(buchner et,al 1998).

3. Background: Similarity, Sequential-data, Sequence and Set Similarity, Distance Measures, Web Usage Data Mining, Cluster Analysis

3.1. Similarity

In many data mining application unlabelled data and we have to group them based on similarity measure(pradeep, Bapi and radhakrishna,2007). Data can be from different application like music files, system calls, transaction records, web logs,genomic data, and so on. In these data there are hidden relations that should be explored to find the interesting information. Formally Similarity is a function S with non negative real values defined on the Cartesian product $X \times X$ of a set X if for every $x, y, z \in X$, the following properties are satisfied by S.

1. Non-Negativity: $S(x, y) \geq 0$.
2. Symmetry: $S(x, y) = S(y, x)$

[1]Associate Professor, Department of CSE,JNTU , Hyderabad.

[2]Assistant professor, Department of CSE, CMEC, Hyderabad

[3]Dept of CSE, bvrit, hyderabad

Email: kakara_2006@yahoo.co.in, devi.duvvuri@ gmail.com, vijaybhaskarbvrit @gmail.com

3. Normalisation : $S(x, y) \leq 1$

4.

3.2. Sequential Data

A sequence is an ordered list of items .A sequence S is denoted as $<s_1, s_2,..s_n>$ where $s_1, s_2, s_3 \ldots s_n$ are called the item sets in the sequence S an item can occur at multiple times in a sequence the number of occurrences of an item in a sequence is called the length of the sequence .A sequence with length l is called the l-sequence the problem of mining sequential patterns was first introduced by Agarwal and Srikanth[1]. In order to find he patterns in the sequences it is necessary to not to look at the items contained in the sequences but also the order of their occurrence.

3.3. Set and Sequence Similarity

A sequence is made up of set of items measure consists of two parts one that quantifies the composition of the sequence and the other that quantifies the sequential nature (sequence similarity).sequence similarity quantifies the amount of similarity in the order of occurrence of item sets within two sequences. Length of the longest common sub sequence (LLCS) with respect to the length of the longest sequence determines the sequence similarity aspect across two sequences

The S^3M, measure for two sequences A and B is given by

$$S^3M, (A, B) = p^* \frac{LLCS(A, B)}{Max(|A|,|B|)} + \frac{q \mid A \cup B \mid}{\mid A \cup B \mid} \qquad (1)$$

Here $p + q = 1$ and $p, q >= 0$. p and q determine the relative weights to be givento the order of occurrence and to the content respectively.

3.4. Disatnce Measures

Euclidean Measure consider two sequences Si, Sj of length n, the Euclidean distance between two points Si and Sj is defined as follows:

$$\text{Euclidean (Si, Sj)} = \sqrt{\sum_{i=1}^{n} (Si - Sj)2} \qquad (4)$$

Jaccards Measure: Jaccard Similarity Measure is defined as the ratio to the number of common item sets and the number of unique item sets in two sequences. Consider two sequences Si and Sj and is defined as

$$\text{Jaccard (Si, Sj)} = \frac{|Si \cap Sj|}{|Si \cup Sj|} \qquad (3)$$

4. CLUSTER ANALYSIS

Clustering is of prime importance in data analysis, machine learning and statistics. Broadly speaking clustering algorithms can be divided into two types partitioned and hierarchical. Partitioning algorithms construct a partition of a database D of n objects into a set of clusters where k is a input parameter.

There are two main issues in clustering techniques. Firstly, finding the optimal number of clusters in a given dataset and secondly, given two sets of clusters, computing relative measure of goodness between them Dbscan is based on two main concepts: density reachability and density connectability. These both concepts depend on two input parameters of the dbscan clustering: the size of epsilon neighborhood e and the minimum points in a cluster m. The number of points parameter impacts detection of outliers. Points are declared to be outliers if there are few other points in the e-Eucledean neighborhood. s.Points are declared to be outliers if there are few of the points in the Neighborhood. DBSCAN starts with a arbitrary point p and retrieves all points density reach ability from p wrt.eps and Minpts. If p is a core point, this procedure yields a cluster w.r.t eps and Minpts. If p is a border point, no points are density reachable from p and DBSCAN visits the next poi of the database. DBSCAN discovers all clusters and detects the noise points. noisy points are generated when a user visits only one web page and such points are not considered to be as a cluster. We extended the dbscan algorithm and named it as seqDbscan. It differs in two aspects. First is sequence representations, and different distance measures.

5. WEB USAGE MINING

The world wide web is a rich source of knowledge that can be used to multi disciplinary applications. Web mining is the use of Data mining techniques to extract information from web documents and services. Web mining is decomposed into three sub tasks.

Web content Mining of text image video metadata and hypertexts to extract useful concepts and rules and summarizes the content no the web.

Web Structure Mining: Mining of underlying link structures of the web in order to categorize web pages measures similarities and reveal relationships between different web sites.

Web Usage Mining: Mining of the data generated by the web users interactions with the web, including web server logs, queries, and mouse clicks in order to extract patterns and trends in web users behavior. Web usage mining focuses on techniques that could predict user behavior while the user interacts with the web.

6. WEB USAGE DATA

In this section we describe the necessary preprocessing steps on the dataset. we also describe here the dataset considered for the experimental results.

Data Preprocessing: Identification of a set of user sessions from the raw usage data provided by the web server. Each user session describes an account of who accessed the web site and what pages are requested and in what order, an for how long each page was viewed. Each user session can be represented as in two ways.

Either as a single transaction of many page references or as a set of many transactions each consisting of a single page reference. In this session we describe MSNBC dataset for experiment.

T1: on-air misc misc misc on-air misc
T2: news sports tech local sports sports
T3: bbs bbs bbs bbs bbs bbs
T4: frontpage frontpage sports news news local
T5: on-air weather weather weather 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

Fig. 1: Example MSNBC Web Navigation Data

Table 1
Number Coding of Web Pages for MSNBC Dataset

Sequences	List
1	6,7,7,7,6,7
2	2,12,3,4,12,12,
3	14,14,14,14,14,14,
4	1,1,12,2,2,4
5	6,8,8,8,12
6	6,6,6,6,3,14
7	1,14,14,1,1,2
8	1,1,1,1,1,14,
9	2,2,15,5,5,16
10	1,11,1,2,2,14

Description of the msnbc Dataset

We collected the data from the UCI dataset repository(http://www.ics.uci.edu) that consists of sever logs from msnbc.com for the month of September 1998. each sequence corresponds to page views of a user during that 24 hour period. Each sequence in the dataset corresponds to the page views of a user during that twenty four r hour period. Each event in the sequence corresponds to a users request for a page.

There are 17 page categories “FrontPage”, “news”, “tech”, “local”, “opinion”, “on-air”, “misc”, “weather”, “health”, “living”, “business”, “sports”, “summary”, “bbs” (bulletin board service), “travel”, “msn-news”, and “msn-sports”.

Each category is associated in order with an integer starting with “1”. For example, “FrontPage” is associated with 1, “news” with 2, and “tech” with 3. Each row below describes the hits in order of a single user. For example, the first

6. Clustering of Web usage Data

In this chapter we use SeqDBSCAN, a standard clustering algorithm and partitions the data in to groups of items that are close to each other based on S^3M measure. SeqDbscan is also used with other measures like Euclidean distance and Jacords similarity measure. In the case of web transactions, each cluster represents a group of transactions that are similar based on the co-occurrence patterns of page categories.

SeqDBSCAN:A New Density Based Spatial Clustering to Sequential Data Description of the SeqDBSCAN Algorithm:

Algorithm 1 : Algorithm for SeqDBSCAN

```
Input:
D = Dataset of N item sets
∈ = Epsilon ranges to [0,1]
Minpts=number of Neighborhood points
Output:
C= Cluster scheme
Begin:
Construct the similarity matrix using S3M measure./
Euclidean distance/jacords
for each unvisited point P in dataset D
mark P as visited
N = get Neighbors (P, T )
if sizeof(N) < MinPts
mark P as NOISE
else
begin
C = next cluster
mark P as visited
end
add P to cluster C
for each point P' in N
if P' is not visited
mark P' as visited
N' = get Neighbors(P', ∈)
if sizeof(N') >= MinPts
N = N joined with N'
if P' is not yet member of any cluster
add P' to cluster C
Step 6: Return C
End
For all clusters
Do
Compute the cluster representatives
End for
```

Consider a data set D ={t1,t2,t3…tN) with N item sets, where each transaction ti=<u1,u2,u3..um> where u2 follows u1 and u3 follows u2 and so on. our objective is to cluster these item sets into distinct clusters .SeqDBSCAN differs from Dbscan in one aspect, i.e. the formulation of the objective function . this algorithm constructs the similarity matrix based on S3M function. Initially the process starts by selecting each sequence from the similarity matrix whose distance is less than the Radius(Epsilon value) as set accordingly in the experiment. Then select each and every unvisited point P, find the number of Neighborhood points of P. If the Neighborhood(P) value is greater than the Minpts, then add it the initial cluster C, similarly consider each and every point P' in N, and repeat the same process, then finally output the cluster C. Then for each cluster a cluster representative is selected. And the inter cluster and the intra cluster similarities are computed based on the Levensthien distance measure. The objective of the SeqDBSCAN algorithm is to find the clusters such that the intra cluster similarity is minimum and the inter cluster similarity is maximum. We used ALD to compute the cost of sequences for each cluster, which represents the goodness of the clusters, ALD is defined as

$$ALD= {}^{1}\sum_{k}^{K}\sum_{j=1}^{|C_j|}\frac{LD(tj,tjs)}{|Cj|} \quad (4)$$

Where k is the total number of clusters, |Cj| is the number of item sets in the jth cluster an LD(tj, tjs) is the levensthein distance between the sth element of the jth cluster to its corresponding cluster representative.

7. EXPERIMENTS USING SEQDBSCAN ON MSNBC DATASET

We considered arbitrary 44,000 web transactions from the MSNBC dataset and performed the experiments and generated clusters using seqDBSCAN clustering technique.

Table 3
Inter Cluster Distance for Clusters Formed from S3M Similarty Matrix

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
C1	0	0.14	0.14	0.15	0.15	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.2	0.21	0.22	0.22	0.21
C2	0.14	0	0.16	0.17	0.17	0.18	0.18	0.19	0.2	0.2	0.21	0.21	0.22	0.22	0.23	0.23	0.24	0.24
C3	0.14	0.16	0	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.2	0.21	0.21	0.22	0.22	0.23	0.23
C4	0.15	0.17	0.16	0	0.18	0.18	0.19	0.19	0.2	0.2	0.21	0.21	0.22	0.22	0.23	0.23	0.24	0.24
C5	0.15	0.17	0.16	0.18	0	0.17	0.18	0.18	0.19	0.19	0.2	0.2	0.21	0.21	0.22	0.22	0.23	0.23
C6	0.16	0.18	0.16	0.18	0.17	0	0.16	0.17	0.17	0.18	0.19	0.19	0.2	0.2	0.21	0.21	0.22	0.22
C7	0.16	0.18	0.17	0.19	0.18	0.16	0	0.21	0.21	0.22	0.22	0.23	0.23	0.24	0.24	0.25	0.25	0.25
C8	0.16	0.19	0.17	0.19	0.18	0.17	0.21	0	0.2	0.2	0.21	0.21	0.22	0.22	0.23	0.23	0.24	0.24
C9	0.17	0.2	0.18	0.2	0.19	0.17	0.21	0.2	0	0.19	0.19	0.18	0.17	0.17	0.16	0.15	0.15	0.15
C10	0.17	0.2	0.18	0.2	0.19	0.18	0.22	0.2	0.19	0	0.14	0.19	0.35	0.44	0.49	0.54	0.74	0.76
C11	0.18	0.21	0.19	0.21	0.2	0.19	0.22	0.21	0.19	0.17	0	0.13	0.13	0.13	0.13	0.12	0.12	0.12
C12	0.18	0.21	0.2	0.21	0.2	0.19	0.23	0.21	0.18	0.16	0.13	0	0.18	0.17	0.17	0.16	0.15	0.15
C13	0.19	0.22	0.21	0.22	0.21	0.2	0.23	0.22	0.17	0.16	0.13	0.18	0	0.37	0.37	0.38	0.38	0.37
C14	0.2	0.22	0.21	0.22	0.21	0.2	0.24	0.22	0.17	0.15	0.13	0.17	0.37	0	0.51	0.56	0.75	0.74
C15	0.21	0.23	0.22	0.23	0.22	0.21	0.24	0.23	0.16	0.15	0.13	0.17	0.37	0.47	0	0.51	0.52	0.51
C16	0.22	0.23	0.22	0.23	0.22	0.21	0.25	0.23	0.15	0.14	0.12	0.16	0.38	0.47	0.51	0	0.57	0.56
C17	0.22	0.24	0.23	0.24	0.23	0.22	0.25	0.24	0.15	0.13	0.12	0.15	0.38	0.48	0.52	0.57	0	0.75
C18	0.21	0.24	0.23	0.24	0.23	0.22	0.25	0.24	0.15	0.13	0.12	0.15	0.37	0.47	0.51	0.56	0.75	0

Table 4
Inter Cluster Distance for Clusters Formed from Euclidean Similarity Matrix

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
C1	0	0.16	0.15	0.15	0.14	0.14	0.14	0.13	0.13	0.12	0.12	0.12	0.11	0.11	0.1
C2	0.16	0	0.16	0.17	0.17	0.18	0.18	0.19	0.2	0.2	0.21	0.21	0.22	0.22	0.23
C3	0.15	0.16	0	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19	0.2	0.21	0.21	0.22
C4	0.15	0.17	0.16	0	0.18	0.18	0.19	0.19	0.2	0.2	0.21	0.21	0.22	0.22	0.23
C5	0.14	0.17	0.16	0.18	0	0.17	0.18	0.18	0.19	0.19	0.2	0.2	0.21	0.21	0.22

Contd...

Contd...

C6	0.14	0.18	0.16	0.18	0.17	0	0.16	0.17	0.17	0.18	0.19	0.19	0.2	0.2	0.21
C7	0.14	0.18	0.17	0.19	0.18	0.16	0	0.21	0.21	0.22	0.22	0.23	0.23	0.24	0.24
C8	0.13	0.19	0.17	0.19	0.18	0.17	0.21	0	0.2	0.2	0.21	0.21	0.22	0.22	0.23
C9	0.13	0.2	0.18	0.2	0.19	0.17	0.21	0.2	0	0.19	0.19	0.18	0.17	0.17	0.16
C10	0.12	0.2	0.18	0.2	0.19	0.18	0.22	0.2	0.19	0	0.17	0.16	0.16	0.15	0.15
C11	0.12	0.21	0.19	0.21	0.2	0.19	0.22	0.21	0.19	0.17	0	0.13	0.13	0.13	0.13
C12	0.12	0.21	0.2	0.21	0.2	0.19	0.23	0.21	0.18	0.16	0.13	0	0.18	0.17	0.17
C13	0.11	0.22	0.21	0.22	0.21	0.2	0.23	0.22	0.17	0.16	0.13	0.18	0	0.37	0.37
C14	0.11	0.22	0.21	0.22	0.21	0.2	0.24	0.22	0.17	0.15	0.13	0.17	0.37	0	0.47
C15	0.1	0.23	0.22	0.23	0.22	0.21	0.24	0.23	0.16	0.15	0.13	0.17	0.37	0.47	0

Table 5
Inter Cluster Distance for Clusters Formed from Jaccard Similarty Matrix on msnbc Dataset

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1	0	0.14	0.14	0.15	0.15	0.16	0.16	0.16	0.17	0.17	0.18	0.18	0.19
C2	0.14	0	0.16	0.17	0.17	0.18	0.18	0.19	0.2	0.2	0.21	0.21	0.22
C3	0.14	0.16	0	0.13	0.14	0.14	0.15	0.17	0.18	0.18	0.19	0.2	0.21
C4	0.15	0.17	0.13	0	0.18	0.18	0.19	0.19	0.2	0.2	0.21	0.21	0.22
C5	0.15	0.17	0.14	0.18	0	0.17	0.18	0.18	0.19	0.19	0.2	0.2	0.21
C6	0.16	0.18	0.14	0.18	0.17	0	0.19	0.19	0.18	0	0.17	0.16	0.16
C7	0.16	0.18	0.15	0.19	0.18	0.19	0	0.21	0.21	0.22	0.22	0.23	0.23
C8	0.16	0.19	0.17	0.19	0.18	0.19	0.21	0	0.2	0.2	0.21	0.21	0.22
C9	0.17	0.2	0.18	0.2	0.19	0.18	0.21	0.2	0	0.19	0.19	0.18	0.17
C10	0.17	0.2	0.18	0.2	0.19	0	0.22	0.2	0.19	0	0.21	0.21	0.22
C11	0.18	0.21	0.19	0.21	0.2	0.17	0.22	0.21	0.19	0.21	0	0.13	0.13
C12	0.18	0.21	0.2	0.21	0.2	0.16	0.23	0.21	0.18	0.21	0.13	0	0.1
C13	0.19	0.22	0.21	0.22	0.21	0.16	0.23	0.22	0.17	0.22	0.13	0.18	0

Table 6
Intra Cluster Distance for Clusters Formed on msnbc Dataset

Intra Cluster	For Table 3	Intra Cluster for Table 4	Intra Cluster for Table 5
C1	0.25.	0.27	0.23.
C2	0.25.	0.27	0.23.
C3	0.25.	0.27	0.23.
C4	0.25.	0.27	0.23.
C5	0.25.	0.27	0.23.
C6	0.25.	0.27	0.23.
C7	0.25.	0.27	0.23.
C8	0.25.	0.27	0.23.
C9	0.25.	0.27	0.23.
C10	0.25.	0.27	0.23.
C11	0.25.	0.27	0.23.
C12	0.25.	0.27	0.23.
C13	0.25.	0.27	0.23.
C14	0.25.	0.27	0.23.
C15	0.25.	0.27	-
C16	0.25.	-	-
C17	0.25.	-	-
C18	0.25.	-	-
ALD	0.19	0.21	0.20

8. CONCLUSIONS

Clustering is an important task in web mining .we considered user sessions comprising of web pages which are sequential in nature. Here in this paper we compared the performance of the SeqDBSCAN algorithm with S3M varying the values of P(sequence similarity) and Q(set similarity),Euclidean, and Jaccards distance measures, over msnbc dataset. Cluster validation is measured using ALD(Average Levensthien distance. From table 6& table 7, ,it is evident that the inter cluster distance among clusters is high in S^3M compared to the Euclidean distance and Jaccards .And the intra cluster distance is mininmum at S^3M then Euclidean and Jaccard., so the user sessions with in the clusters formed based on S^3M have retained more sequential information when compared to other two measures. This results shows that the S^3M measure concentrates not only on the content of the data, but on the order of occurrence of data.

REFERENCES

[1] Giedre Grizaite, Roland Innerhofer-Oberperfler(2005), DBSCAN Clustering Algorithm.

[2] Martin Ester, Hans-Peter Kriegel, Jorg sander, Xiaowei Xu. A Density–Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Munchen, Germany.

[3] Agarwal and Srikanth(1994). "Mining Sequential Patterns. Proceedings of the International Conference on Data Engineering (pp, 3-14)", IEEE Computer Society Press, Taipie, Taiwan.

[4] Bergroth, L, Hakonene, H, & Raita,T.(2000). "A Survey of Longest Common Subsequence Algorithm". The 7th International Symposium on String Processing and Information Retrieval (pp 39-48).

[5] Cooley R, Mobasher, B., & Srivastava, J.(1999). "Data Preparation for Mining World Wide Web Browsing Patterns", Knowledge and Information Systems, 1(1), 5-32.

[6] Dunham. H.(2003). Data Mining : Introductory and Advanced Topics: Prentice Hall.

[7] Jain., A.K, Murthy, M.N & Flynn, P.J.(1999). "Data Clustering: A Review", ACM Computing Surveys, 31(3), 264-323.

[8] Shahabi, c ., zarkesh, A.M, Adibi, J&Shah, V.(1997). "Knowledge Discovery from Users Web Page Navigation Data". Proceedings of the 7th International Workshop on Research Issues in Data Engineering High Performance Database Management for Large Scale Application, (p.20), IEEE Computer Society, Washington, DC, USA.

[9] Tan Taniar and & Smith, K.A (2005). "A Clustering Algorithm based on an Estimated Distribution Model". Interbational Journal of Business Intelligence and Data Mining. 1(2), 229-245, Inderscience Publishers.

[10] Mobhasher, B.(2004). "Web usage Mining and Personalization". CRC Press.

[11] Mobhasher,B. Cooley, R., & Srivatava, J.(1999). "Creating Adaptive Web Sites through Usage based Clustering of URLS", Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange.