Web Usage Data Clustering using Dbscan algorithm and Set similarities

K.Santhisree
Associate Professor
CSE, JNTU
Hyderabad,India
Kakara_2006@yahoo.co.in

Dr A. Damodaram
Director of UGC-ASC
CSE, JNTU
Hyderabad, India
damodarama @ rediffmail.com

S.Appaji
Assiistant professor
Arora Engg college,
Hyderabad, India
appaji_sv@yahoo.co.in

D.NagarjunaDevi
Assistant Professor
CSE, CMEC
Hyderabad,India
devi.duvvuri@gmail.com

Abstract: Web usage mining is the application of data mining techniques to web log data repositories. It is used in finding the user access patterns from web access log. User page visits are sequential in nature. In this paper we presented new Rough set Dbscan clustering algorithm which identifies the behavior of the users page visits, order of occurrence of visits . Web data Clusters are formed using the rough set Similarity Upper Approximations. We present the experimental results on MSNBC web navigation dataset, and proved that Rough set Dbscan clustering has better efficiency and performance clustering in web usage mining is finding the groups which share common interests compared to Rough set agglomerative clustering .

Keywords: Sequence, Web usage data, set approximations, set similarity, rough sets

1.Introduction:

Clustering is of prime importance in data analysis, machine learning and statistics. It is defines as the process of grouping N item sets into distinct clusters based on similarity or distance function A good clustering technique may yield clusters thus have high inter cluster and low intra cluster distance. (Pradeep Kumar and bapi, 2007)The objective of clustering is to maximize the similarity of the data points within each cluster and maximize dissimilarity across clusters.

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. Hierarchical algorithms create decomposition of the database D. they are a Agglomerative and divisive. Hierarchical clustering builds a tree of clusters, also known as a dendrogram. Every cluster node contains child cluster. An agglomerative clustering starts with one-point (singleton)

In conventional clustering objects that are similar are allocated to the same cluster while objects that differ are put in different clusters. These clusters are hard clusters.. In soft clustering an object may be in more than two or more clusters.

A rough clustering is defined in a similar manner to a rough set . the lower approximation of a rough cluster contains objects that only belong to that cluster. The upper approximation of a rough cluster contains objects in the clusters which are also members of other clusters.

2. Literature Review:

2.1. Sequential data: A sequence is an ordered list of items .A sequence S is denoted as <s1,s2,..sn> where s1,s2,s3…sn 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 an Srikanth). 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. In order to find patterns in sequences , it is necessary to not only look at

978-0-7695-3958-4/10 $26.00 © 2010 IEEE
DOI 10.1109/DSDE.2010.14

the items contained but also on the order of their occurrences.

2.2 Set Similarity, sequence similarity : In data mining applications, we are given with unlabelled data (IJDWM 2007 pradeep,Bapi)) and we have to group them based on the similarity measure. These data may arise from diverse application domains. They may be music files, system calls, transaction records, Web logs, and soon. In these data, there are hidden relations that should be explored to find interesting information. For example, from Web logs, one can extract the information regarding the most frequent access path, One can extract features from sequential data to quantify parameters expressing similarity. Similarity is a function S with nonnegative real values defined on the Cartesian product $X \times X$ of a set X. It is called a metric on X if for every *x, y, z, the* following properties has to be satisfied by S.

1. **Non-negativity:** $S(x, y) \geq 0$.
2. **Symmetry:** $S(x, y) = S(y, x)$.
3. **Normalization:** $S(x, y) \leq 1$.

. A new measure called the sequence and the set similarity measure S3M introduced(pradeep kumar) which consists of two parts. One that quantifies the composition of the sequence(set similarity) and the other that quantifies the sequential nature(sequence similarity) sequence similarity quantifies the similarity on the order of occurrence of item sets within two sequences. Length of the longest common subsequence (LLCS) with respect to the length of the longest sequence determines the sequence similarity aspect across two sequences .consider two sequences X and Y the sequence similarity is measured as(Pradeep and P.R krishna,2007)

- $\text{SeqSim}(X,Y) = \dfrac{\text{LLCS}(X,Y)}{\text{Max}(|X|,|Y|)}$
- $\text{SetSim}(X,Y) = \dfrac{|X \cap Y|}{|XUY|}$

Set similarity is defined as the ratio to the number of common item sets and the number of unique item sets in two sequences, thus for two sequences X,Y the set similarity is defined as Thus, S3M measure for two sequences A and B is given by:

$$S3M(X,Y) = P^* \frac{\text{LLCS}(X,Y)}{\text{Max}(|X|,|Y|)} + q^* \frac{|X \cap Y|}{|XUY|}$$

Here, $p + q = 1$ and $p, q \geq 0$. p and q determine the relative weights to be given for order of occurrence (sequence similarity) and to content (set similarity),respectively. In practical a LLCS between two sequences can be found by the programming approach

3. Web usage mining: 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.
Resource finding,Generalisation and Analysis.
web content mining of text image video metadata and hypertexts to extract useful concepts and rules and summarizes the content on 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 I 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. The web usage mining process could be classified into two commonly used approaches. One important use of clustering in web usage mining is finding the groups which share common interests and behavior by analyzing the data collected in the web servers. This study contributes the topic clustering of web usage data and shows the interests and behaviors of the various user visits

4.Algorithm: Proposed new Rough set Dbscan clustering:

Input : N
T: A set of n transactions ϵ U
Threshold d ϵ [0,1]
Relative similarity r ϵ [0 1]
Epsilon ϵ[0,1]

Minpts:Number of Neighbor hood points
Output:
Cluster scheme C
Begin
Step 1: Construct the similarity matrix using S3M measure(Definition 1).
Step 2: select all points from D that satisfy the Eps and Minpts
C = 0
for each unvisited point P in dataset D
mark P as visited
N = get Neighbors (P, eps)
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' = getNeighbors(P', eps)
if sizeof(N') >= MinPts
N = N joined with N'
if P' is not yet member of any cluster
add P' to cluster C
Step 3:
Return C
Step 4:
For all Ti ϵ U Compute Si= R(Ti)
Using definition 2 for given threshold d.
Step 5:
Next compute the constrained-similarity upper Approximations Sj for relative similarity r using definition 3
if Si= SJ
end if
Step 6.: Repeat step 3 until U≠ Ø;
Return D.
End

5.Experimental results:

Description of the Dataset
We collected the data from the UCI dataset repository(http://www.ics.uci.edu) that consists of several 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 "% Sequences:" describes the hits in order of a single user.

Table1: sequence dataset

sequence	Order of user page visits
1	1 1
2	2
3	3 2 2 4 2 2 2 3 3
4	5
5	1
6	6
7	1 1
8	6
9	6 7 7 7 6 6 8 8 8 8
10	6 9 4 4 4 10 3 10 5 10 4 4 4
11	1 1 1 11 1 1 1
12	12 12
13	1 1

Step1: To illustrate our example we considered 13 data sequences, AND similarity table was computed using similarity metric with p=0.5.Refer to the samples at the above mentioned website.
Step2:The first similarity upper approximation at threshold value 0.5 is given by
R(T1)={T1,T5,T7,T11,T13},
R(T2)={T2},
R(T3)={T3},R(T4)={T4,}R(T5)={T1,T5,T11,T13}
R(T6)={T6,T8}

R(T7)={T1,T7,T11,T13}
R(T8)={T6,T8}
R(T9)={T9},R(T10)={T10}
R(T11)={T1,T5,T7,T11,T13}
R(T12)={T12}
R(T13)={T1,T5,T7,T11,T13}

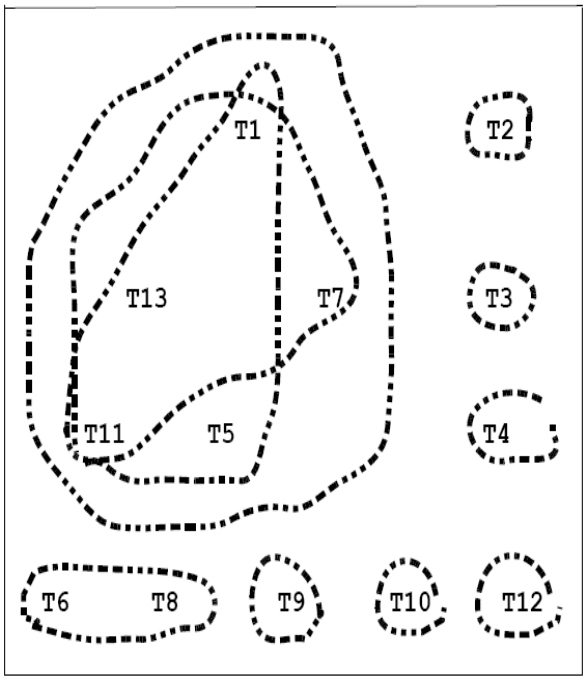


Figure 1(first upper Approximations)

The second upper approximations with the proposed value 0.5 are
RR(T1)={T1,T5,T7,T11,T13},
RR(T2)={T2,},
RR(T3)={T3},RR(T4)={T4}
RR(T5)={T1,T5,T7,T11,T13},
RR(T6)={T6,T8},
RR(T7)={T1,T5,T7,T11,T13}
RR(T8)={T6,T8},RR(T9)={T9,},
RR(T10)={T10},
RR(T11)={T1,T5,T7,T11,T13},
RR(T12)={T12}
,RR(T13)={T1,T5,T7,T11,T13}

STEP 3:

The third upper approximations
RRR(T1)={ T1,T5,T7,T11,T13},
RRR(T2)={T2,}
RRR(T3)={T3},
RRR(T4)={T4},
RRR(T5)={ T1,T5,T7,T11,T13},
RRR(T6)={T6,T8}
RRR(T7)={T1,T5,T7,T11,T13}
RRR(T8)={T6,T8},
RRR(T9)={T9,},
RRR(T10)={T10}
RRR(T11)={T1,T5,T7,T11,T13},
RRR(T12)={T12}
RRR(T13)={T1,T5,T7,T11,T13}.

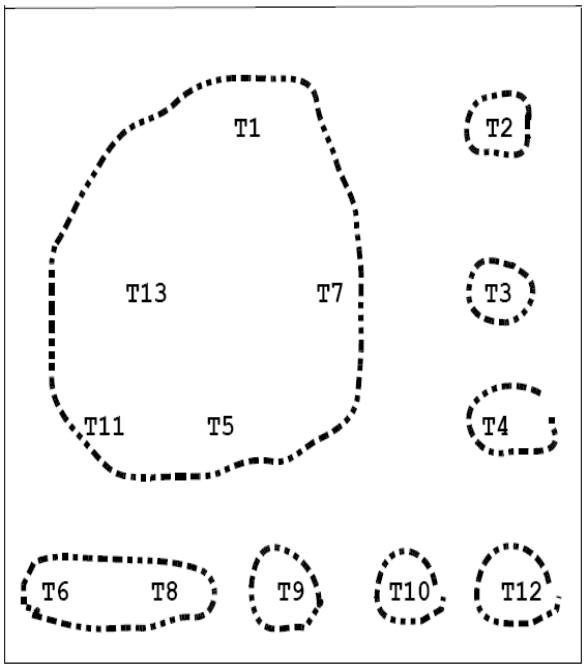
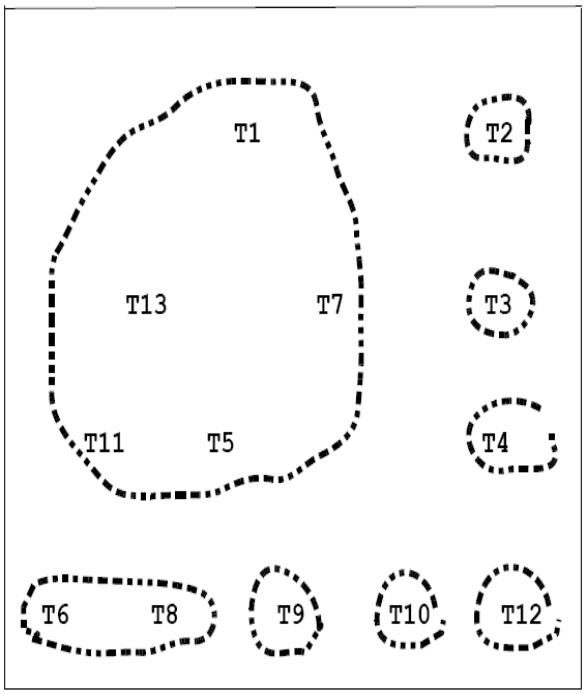


Figure 2(second upper Approximations)



.Figure3: (Third upper Approximations)

The similarity upper approximations for C1={ T1,T5,T7,T11,T13}
C2={T2,},C3={T3},C4={T4},C5={T1,T5,T7,T11,T13},C6={T6,T8},C7={T1,T5,T7,T11,T13},C8={T6,T8},C9={T9,},C10={T10},C11={T1,T

The clusters formed after applying Dbscan clustering algorithm on the 13 sets with Epsilon value=0.2 and Epsilon neighbor is 3
C1={T5,T7,T13},C2={T2}C3={T3},C4={T4},C5={T1},C6={T8},C7={T7},C8={T6},C9={T9

6.Comparisionofexperimentalresults

The clusters formed using rough set agglomerative clustering are C1={T1,T5,T7,T11,T13},C2={T2,},C3={T3},C4={T4},C5={T1,T5,T7,T11,T13},C6={T6,T8},C7={T1,T5,T7,T11,T13},C8={T6,T8},C9={T9,},C10={T10},C11={T1,T5,T7,T11,T13},C12={T12},C13={T1,T5,T7,T11,T13}.

The clusters formed after applying Dbscan clustering algorithm on the 13 sets (with Epsilon value=0.2,epsilonneighbour=3),areC1={T5,T7,T13},C2={T2}C3={T3},C4={T4},C5={T1},C6={T8},C7={T7},C8={T6},C9={T9},C10={T10},C11={T11}} C12={T12},C13={T1}.T1 is a border point which results in the clusters c13 and c5.

7.Conclusions:

In this paper we developed a new rough set dbscan clustering algorithm and presented a 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. The rough set dbscan clustering algorithm is efficient when compared to the rough set agglomerative clustering . As in rough set agglomerative clustering the elements can be present in more than one cluster(soft clustering) , where as in our proposed algorithm rough set dbscan algorithm(hard clustering), the elements will not occur in other clusters

5,T7,T11,T13},C12={T12},C13={T1,T5,T7,T11,T13}.
STEP4: C=0
STEP5:

},C10={T10},C11={T11}}
,C12={T12},C13={T1}
T1 is a border point which results in the clusters c13 and c5.

References:

[1] Supriya kumar and P.Radhakrishna, clustering of sequential data. using rough sets, Journal of DKE , pp[183-199],2007
[2]Raymond kosala, Hendricks blockeel: web usage mining : A survey.ACM SIGKDD Explorations 2, issue 1(june 2000)pp[1-15]
[3]pradeep kumar and P.R Krishna, raju s.bapi :SeqPam:a sequence clustering algorithm for web Personalization.IJDWM 3(1),],P [29-53],2007
[4] supriya kumar , P.radhakrishna clustering webtransactions using rough approximation.fuzzy sets and systems,Elsevier,,pp [131-138],2004
[5]Pavel Berkhin accrue software,inc Asurvey of clustering algorithms, proceedings of the 29th VLDB conference ,2003..[9].