Comparison of Clustering Techniques: PAM and SSM-PAM: Experiments and Test cases

Dr.K.Santhi Sree
Professor of CSE
JawaharLal Nehru Technological University
Kukatpally, Hyderabad
kakara_2006@jntuh.ac.in

***Abstract*--**Clustering web usage data is useful to discover interesting sequential patterns related to user traversals, behavior and their characteristics, which helps for the improvement of better Search Engines and 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. The other issue that arises is how to measure similarity between sequences. There exist multiple similarity measures in the past like Euclidean , Jaccard ,Cosine and many. Most of the similarity measures presented in the history deal only with sequence data but not the order of occurrence of data. A novel similarity measure named SSM(Sequence Similarity Measure) is used that shows the impact of clustering process ,when both sequenc**e** and content information is incorporated while computing similarity between sequences. SSM (**S**equence **S**imilarity measure) captures both the order of occurrence of page visits and the page information as well ,and compared the results with Euclidean, Jaccard and Cosine similarity measures. Incorporating a new similarity measure, the existing PAM algorithms are enhanced and the new named as SSM-PAM for Web personalization. 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 similarity measure has significant results when comparing similarities between web sessions with other previous measures , and provided good time requirements of the newly developed SSM-PAM algorithms. Experiments are performed on MSNBC.COM website (free online news channel), in the context of Partitioning clustering in the domain of Web usage mining.

Keywords*-Data Mining, Clustering, Similarity measures, Web Personalization, PAM and SSM-Kmeans and SSM-Kmedoids., Sequence Mining, Clustering .Partitioning algorithms.*

____*****____

1. INTRODUCTION

A. Data Mining

Data mining, called Knowledge Discovery in Databases (KDD) an interdisciplinary subfield of computer science is the process of identifying knowledge / patterns in large heterogeneous data sets .The goal of the Data mining process is to extract information from a data set, preprocess and transform it into an understandable structure for further use. Various stages of Data mining are Selection, Preprocessing, Transformation, Data mining, Interpretation and evaluation. The various Data mining techniques are Classification, Clustering, Prediction, Association and Discrimination.

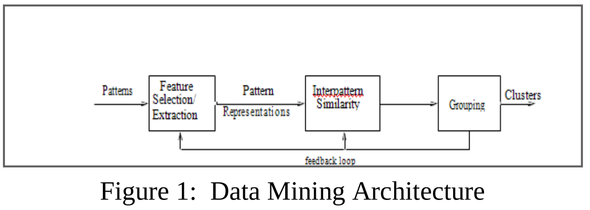

Figure 1: Data Mining Architecture

B.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. The most popular clustering techniques are PAM algorithms (k-Means) and K-Medoids) . In cluster analysis, the *k*-means algorithm can be used to partition the input data set into *k* partitions (clusters). PAM algorithms a finds clusters only in the spherical shape where as Density based clustering techniques finds clusters of arbitrary shape.

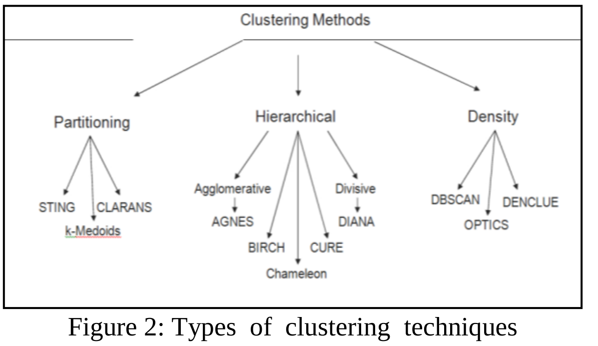

Figure 2: Types of clustering techniques

C. Sequence Mining

Sequential Pattern mining is interdisciplinary subfield of Data mining concerned with finding relevant patterns described in a sequence. Given a sequence database D={S1,S2,...,Sn} where each sequence S is an ordered list

 ISSN: 2321-8169
1465 – 1470

of events/items <i1, i2,...,in>.There are several key traditional computational problems addressed within this field. These include building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequences for similarity, and recovering missing sequence numbers. In general, sequence mining problems can be classified as string mining which is typically based on string based algorithms and itemset mining which is typically based on association rule mining.

D. Web Personalization

Web personalization is the process of identifying what users are exactly looking for on the web, their traversals and their behavior. Due to the continuous growth of the Web data, Web personalization has become one of the challenging task for the researchers and commercial areas. The steps of a Web personalization process include: the collection of Web data, modeling and categorization of these data (preprocessing phase), the analysis of the collected data, the determination of the actions that should be performed. Web data are collected and used in the context of Web personalization. These data are classified in four categories .web Structure data represent how pages are linked to one another. Web usage data represents what users are exactly looking for on the Web and their characteristics such as a visitor's IP address, time and date of access, complete path (files or directories) accessed, referrers' address, and other attributes that can be included in a Web access log.

E. 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 vector based. Euclidean distance measure is frequency based similarity measures for two sequences S_1 and S_2 in an N-dimensional space. It is defined as the square root of the sum of the corresponding dimensions of the vector. 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

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

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.

$$(Sim(S_1,S_2)) = \frac{S_1S_2}{|S_1|^2+|S_2|^2-S_1S_2}$$

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

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

F. SSM-Sequence Similarity Measure

In this work a novel similarity measure **[2]** is used that captures both the order of information as well as content(information) called the SSM(sequence similarity measure).

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

2. EXISTING METHODOLOGY

Usually when dealing with sequences, the data is converted into n-dimensional frequency vectors. The vector representation can be either indicating presence or absence of symbol in a sequence, or, indicating frequency of symbol within a sequence. While computing similarity between sequences they either consider the content /information or the order information. 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, Jaccard, Cosine. PAM algorithms can be applied for clustering. Similarity are calculated which illustrates the similarity between the sequences. And the Inter cluster similarity has to be maximized and Intra cluster similarity has to be minimized.

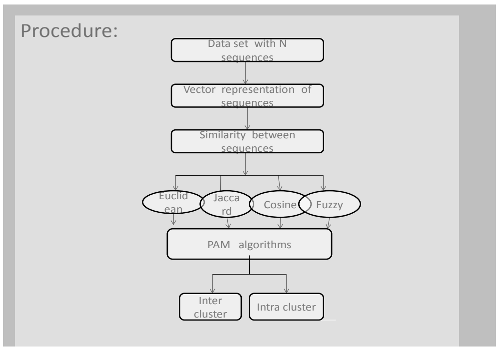

Figure 3: Exiting Work Procedure

3. PROPOSED WORK

 ISSN: 2321-8169
1465 – 1470

The work concentrates on Clustering technique on the domain of web usage data. A new similarity measure [6] is used to measure similarity/distance between two sequences and experiments are conducted on PAM algorithms..In all the experiments the running time of the new similarity measure is accurate and best compared to the earlier similarity measures. An experimental framework for sequential data stream mining on clustering on web usage data is built.

A. Experimental Results

a)Web Navigation dataset used for Testing

MSNBC is a joint venture between Microsoft and NBC(National Broad casting) is a famous online news website with has different news subjects. There are 17 categories of news like frontpage,news,tech,local,opinion,onair,weather,health,living,business,sports,summary,bbs,travelmisc,msn-news and msn-sports. For example, 'frontpage' is coded as 1, 'news' as 2, 'tech' as 3, etc. Web Navigational dataset is considered in Table 5.1

Table 1. Web Navigational Dataset

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

Table 2. Converted Web Navigational dataset

Sequence	Order of page visits
T1	6,15,15,15,6,15
T2	2,11,3,4,11,11
T3	11,13,13,13,13,13
T4	1,1,11,2,2,4
T5	6,7,7,7,11,11
T6	6,6,6,6,3,6
T7	1,13,13,1,1,2
T8	1,1,1,1,1,13
T9	2,2,14,5,5,16
T10	1,10,1,2,2,13

B.PAM and SSM-PAM Clustering Technique

a) Experiments on Synthetic web Navigational Dataset for PAM algorithms.

Consider arbitrarily 100 records of web transactions from MSNBC.COM website. The transactions are converted to vector representation, and a 100 X 100 similarity matrix is computed using Euclidean distance measures mentioned above. In the step two after applying existing K-Means clustering technique the clusters formed are 08. Table 8 × 8 matrix which shows the Inter cluster distance using Euclidean distance measure. For example, the Inter cluster distance (C1,C2) =0.15. and Inter cluster distance between the clusters (C3,C8)=0.15. That is the patterns lying in the clusters C1,C2,C3,C8 are more similar when compared to the patterns lying in the other clusters.

Table 3: Inter Cluster Distance Using Euclidean Distance Measure for K-Means

Ci XCi	C1	C2	C3	C4	C5	C6	C7	C8
C1	-	0.15	0.16	0.16	0.16	0.16	0.17	0.17
C2	0.15	-	0.13	0.13	0.14	0.13	0.13	0.14
C3	0.16	0.13	-	0.12	0.14	0.15	0.15	0.15
C4	0.16	0.13	0.12	-	0.18	0.18	0.18	0.19
C5	0.16	0.14	0.14	0.18	-	0.16	0.16	0.16
C6	0.16	0.13	0.15	0.18	0.16	-	0.16	0.17
C7	0.17	0.13	0.15	0.18	0.16	0.16	-	0.21
C8	0.17	0.14	0.15	0.19	0.16	0.17	0.21	-

Table 4: Inter Cluster Distance Using Jaccard Distance Measure For K-Means

Jaccard	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	-	0.15	0.16	0.16	0.16	0.16	0.17	0.17	0.18
C2	0.15	-	0.13	0.13	0.14	0.13	0.13	0.14	0.15
C3	0.16	0.13	-	0.12	0.14	0.15	0.15	0.15	0.16
C4	0.16	0.13	0.12	-	0.18	0.18	0.18	0.19	0.19
C5	0.16	0.14	0.14	0.18	-	0.16	0.16	0.16	0.17
C6	0.1	0.1	0.1	0.1	0.1	-	0.1	0.1	0.1

 ISSN: 2321-8169
1465 – 1470

	6	3	5	8	6		6	7	8
C7	0.17	0.13	0.15	0.18	0.16	0.16	-	0.21	0.21
C8	0.17	0.14	0.15	0.19	0.16	0.17	0.21	-	0.18
C9	0.18	0.15	0.16	0.19	0.17	0.18	0.21	0.18	-

Table 5: Inter Cluster Distance Using Cosine Similarity Measure for K-Means

Cosine	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	-	0.16	0.17	0.17	0.17	0.18	0.19	0.21	0.19
C2	0.16	-	0.17	0.18	0.17	0.17	0.18	0.18	0.17
C3	0.17	0.17	-	0.11	0.12	0.13	0.14	0.14	0.14
C4	0.17	0.18	0.11	-	0.16	0.13	0.19	0.17	0.17
C5	0.17	0.17	0.12	0.16	-	0.13	0.2	0.18	0.18
C6	0.18	0.17	0.13	0.13	0.13	-	0.21	0.18	0.18
C7	0.19	0.18	0.14	0.19	0.2	0.21	-	0.18	0.22
C8	0.21	0.18	0.14	0.17	0.18	0.18	0.18	-	0.23
C9	0.19	0.17	0.14	0.17	0.18	0.18	0.22	0.23	-

b) Experiments on Synthetic web Navigational Dataset for Kmedoids

Consider arbitrarily 100 records of web transactions from MSNBC.COM website. The transactions are converted to vector representation, and a 100 X 100 similarity matrix is computed using Euclidena measure mentioned above. In the step two after applying K-medoids clustering technique the clusters formed are 11 .Table 6 11 X 11 matrix which shows the inter cluster distance using Euclidean distance measure.

Table 6: Inter Cluster Distance Using Euclidean Distance for Kmedoids

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
C1	-	0.16	0.17	0.17	0.17	0.18	0.19	0.21	0.19	0.19	0.19
C2	0.16	-	0.17	0.18	0.17	0.17	0.18	0.18	0.17	0.16	0.16
C3	0.17	0.17	-	0.11	0.12	0.13	0.14	0.14	0.14	0.13	0.14
C4	0.17	0.18	0.11	-	0.11	0.12	0.14	0.16	0.18	0.12	0.14
C5	0.17	0.17	0.12	0.11	-	0.17	0.16	0.17	0.17	0.16	0.16
C6	0.18	0.17	0.13	0.12	0.17	-	0.11	0.12	0.12	0.13	0.14
C7	0.19	0.18	0.14	0.14	0.16	0.11	-	0.19	0.18	0.19	0.17
C8	0.21	0.18	0.14	0.16	0.17	0.12	0.19	-	0.17	0.16	0.21
C9	0.19	0.17	0.14	0.18	0.17	0.12	0.18	0.17	-	0.21	0.22
C10	0.19	0.16	0.13	0.12	0.16	0.13	0.19	0.16	0.21	-	0.17
C11	0.19	0.16	0.14	0.14	0.16	0.14	0.17	0.21	0.22	0.17	-

c) Experiments on Standard web Navigational Dataset.

Considered transactions of varying sizes of 5000, 10000,20,000,30000,40000 from MSNBC dataset. Table 7 shows the number of clusters formed by applying the existing PAM clustering technique and enhanced SSM-PAM. The Inter cluster similarity and Intra cluster similarity are calculated. That demonstrates the useful ness of sequential mining in the domain of web usage data .

Table 7. Inter and Intra cluster distance for PAM algorithms

PAM(K-Means) 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	0.054	0.039	0.033	0.059	0.036
Average Intra cluster	4.27	4.000	4.989	6.867	5.896
PAM-(K-Medoids) CLUSTERING RESULTS USING EUCLIDEAN					
No of samples	5000	10000	20000	30000	40000

ISSN: 2321-8169
1465 – 1470

No of clusters formed	99	114	147	135	197
Inter cluster	4.281	4.317	5.213	8.153	7.298
Average Inter cluster	0.043	0.037	0.035	0.045	0.026
Average Intra cluster	4.013	4.291	5.222	7.293	8.123

Table 7. Inter and Intra cluster distance for SSM-PAM algorithms

(SSM-KMEANS) CLUSTERING RESULTS USING SSM					
No of samples	5000	10000	20000	30000	40000
No of clusters formed	93	113	124	138	174
Inter cluster	4.6	4.8	5.39	7.123	7.932
Average Inter cluster	0.049	0.042	0.043	0.044	0.045
Average Intra cluster	4.001	4.019	4.318	5.293	6.142
(SSM-KMEDOIDS) CLUSTERING RESULTS USING SSM					
Size of sequences	5000	10000	20000	30000	40,000
No of clusters	94	126	149	141	187
Inter cluster	4.69	4.47	5.213	6.153	7.298
Average Inter cluster	0.049	0.035	0.035	0.043	0.039
Average Intra cluster	3.314	3.187	4.212	3.297	4.123

4. TIME REQUIREMENTS

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

Table 8 Time Requirements of PAM and SSM-PAM

PAM-KMEANS					
Size of sequences	5000	10000	20000	30000	40,000
No of clusters	83	124	155	116	189
Time taken in seconds	1566	2665	2785	3218	3196
PAM-KMEDOIDS					
Size of sequences	5000	10000	20000	30000	40,000
No of clusters	99	114	147	135	197
Time taken in seconds	1624	2660	2794	3301	3126
SSM-KMEANS					
Size of sequences	5000	10000	20000	30000	40,000
No of clusters	93	113	124	138	174
Time taken in seconds	1085	1879	3643	1956	2498
SSM-KMEDOIDS					
Size of sequences	5000	10000	20000	30000	40,000
No of	94	126	149	141	187

clusters					
Time taken in seconds	1080	1871	2946	1749	2156

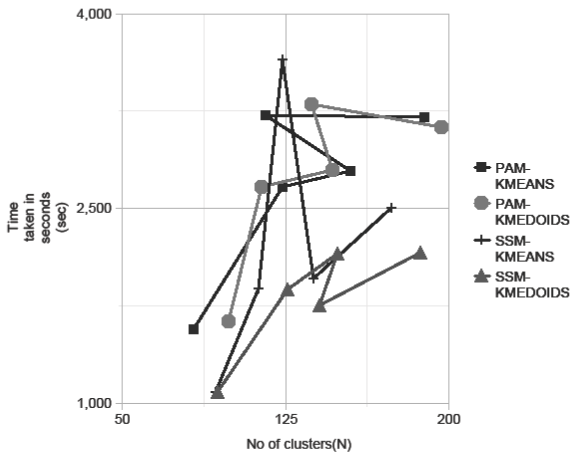

Figure 4: Time taken for K-Means ,K-mediods, SSM-Kmeans, SSM-Kmedoids

5. CONCLUSIONS

Considered arbitrarily web transactions of equal length from the MSNBC dataset and performed the experiments PAM and SSM-PAM clustering techniques. We used previously existing four different distance/similarity measures namely Euclidean , Jaccard, Cosine, and the newly developed measure called SSM. In PAM the number of clusters are 08,10,09 respectively . For good clustering algorithm, the intra cluster distance should be minimum. SSM measure which is a combination of sequence as well as set measure, confirms that the web clustering should consider the sequence as well as content value. For example in SSM-PAM for 5000 samples ,the time taken for execution are 1085,1879,3643,1956,2498 respectively. The time taken to execute the algorithm SSM-PAM is less when compare to other previous PAM clustering techniques .

Experiments are performed in the context of Partitioning clustering. A new similarity measure for sequential data (*SSM*) is devised and used and incorporated SSM with PAM for Web Usage sequential data. Our results by explanations and conclusions, finally showed behavior of clusters that made by enhanced SSM-PAM clustering techniques on a sequential data in a web usage domain. This new SSM-PAM required less time complexity then the existing.. 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.

a) Future Work

we extend our work in future to other clustering techniques and to other domains as well.

- Developing new similarity measures for continuous and discrete sequential data.
- Applying these new clustering techniques to the domains like medical, defense, bioinformatics etc.
- This work can be extended to sequences of unequal length.
- The time complexities of the proposed algorithms can be improved further.

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

[3]. 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.

[4]. Han.J, Kamber.M, “Data Mining Concepts and Techniques, Morgan Kaufmann Publishers”, cluster analysis, pp.(339-.352), 2001.

[5]. 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.