A Novel Framework for Geo-clustering of User Movements based on Trajectory Data

V. Tanuja and P. Govindarajulu
Sri Venkateswara University,
Tirupati, AP india

Abstract
Data mining applications on real time data sources is an evolving area of research .The availability of data gathering techniques in modern era is facilitating novel means to go with extracting knowledge from new data trends. Trajectory data is one of the potential sources to study the style of the data being generated through sensing the events with special and temporal nature. Trajectory data clustering is a major data mining technique to deal with the trajectory data of real-time movements. In this paper a new algorithm is proposed to cluster the trajectory data. The data regarding the sequential locations of user movements is considered for clustering. The trajectory data of taxy/auto riksha movements in the Nellore town of Andhra Pradesh is considered for evaluating the algorithm. The results showed that the proposed algorithm is better than the existing algorithms in the literature in terms of computational effort needed. This novel approach will met with the process of knowing user profile behavior and guide towards profitable Customer Relationship Management (CRM) practices in multifold business situations.

1. Introduction:

The trend of data gathering techniques have been changed from manual and semi-automated to full automated sensing systems. The seamless availability of data collection systems from the practical situations, and the variety of data available with higher volumes is a real challenge to deal with the same to use the data .Trajectory data resembles the sequence data with an order of user movements. The study of user movement data needs the grouping of such data into partitions where each partition represent specific behavior. To obtain such groups means are required to compare the data objects.

1.1Trajectory Data:

Trajectory data refers to the data representing a sequence of special or temporal or eventual data. Trajectory data represents information that describes the movement of the user in time and space. It is represented by a sequence of time stamped geographical location. Nowadays, there is a tremendous increase of moving objects databases due to on the one hand, location acquisition technologies land like GPS and GSM networks and, on the other hand computer vision based tracking techniques [5]. Two main challenges must be considered while dealing with trajectory data – first, how should one can extract each group of moving objects that having the similar moving styles?. This can be solved by applying a clustering technique on the trajectories because it groups similar ones while separating dissimilar ones. Second an important characteristic of a trajectory is that its movement is not uniform in both the spatial and temporal domains [11]. Huey-Ru Wu et al. [11] proposed DivCluST, an approach to finding regional typical moving styles by dividing and clustering the trajectories in consideration of both the spatial and temporal constraints.

Costas Panagiotakis et al. [5] proposed an index based global voting method that allows a technique to represent the representativeness of a trajectory in a moving object databases (MOD) as a smooth continuous descriptor and authors also introduced an algorithm for the automatic segmentation of trajectories into homogeneous sub trajectories according to their representativeness in the MOD. C. Panagiotakis et al. [6] proposed an approach for expressing the “representativeness” in MOD via a voting process that is applied for each segment of a given trajectory.

Gook-Pil Roh et al. [6] proposed a framework for supporting effective queries over trajectories, authors have identified five desirable properties – Road network, Spatial proximity, Whole Pattern Matching, Sub pattern matching, Reverse sub pattern matching, that have not been fully satisfied by any of the prior work. Aiden Nibali et al. [1] developed an efficient trajectory compression technique called TRAJIC with linear run-time complexity (O(N)).

These are the foundation of many applications such as traffic analysis, mobile user’s behavior, market analysis, store keeping and many more. Trajectories provide knowledge to estimate, compare and construct candidate routes by historical data objects. Delta compression achieves lossless compression by storing the difference between successive data points in a trajectory rather than the points themselves [1].

Manuscript received March 5, 2017
Manuscript revised March 20, 2017

Moving objects can be human being, animals, vehicles and even natural phenomenon. Destination prediction technique involves in tracking the trace of the moving objects with the help of positioning services like GPS equipped device. Trajectories provide intelligence to estimate, compare and construct candidate routes by historical road network [3].

1.2 Trajectory Data Patterns

Considerable effort has been devoted to discovery of trajectory patterns in data mining and computational geometry [12]. Jae-Gil Lee et al. [12] proposed a unifying framework of mining trajectory patterns of various temporal tightness. Trajectory patterns can be classified mainly into three types depending on the tightness of temporal constraints – 1) Flock Patterns, Convoy Patterns or Moving Clusters, 2) Time relaxed Trajectory Joins or Hot Motion Paths, 3) Sub-Trajectory Clusters [12]. Jae-Gil Lee et al. [13] proposed a framework of frequent pattern based classification for trajectories on road networks.

The increase in availability of location acquisition technologies such as GPS set on cars, WLAN networks, and mobile phones carried by people have enabled tracking almost any kind of moving objects, which results in huge volumes of spatio-temporal data in the form of trajectories [16].The recent proliferation of ubiquitous sensing technologies, intelligent transportation systems, and location based services increases the availability of human trajectories [17].Trajectory data reflects human mobility, it can be naturally utilized for location based recommendation applications, including personalized location prediction, group based location recommendation, and user mobility modeling [18]. Trajectory pattern mining aims to discover frequent sequential patterns that are sequential relationships among regions [22].

1.3 Trajectory Data Clustering:

Clustering of users is one of the most important applications in finding movement based communities based on sequential movement details of users. Users in the same movement group have similar features whereas users in the different movement groups have dissimilar features. Forming movement based communities is very useful in many real time applications such as trajectory recommendation services and location based services. Hnin Su Khaing et al. [8] proposed an efficient clustering algorithm which can solve the problem of DBSCAN for moving object trajectories and this algorithm consists of three phases – portioning, clustering and grouping. Movement based communities are created based on the similarity relationship between users. Trajectory data mining focuses more on the mining of trajectories associated with geometric shapes such as clustering and finding outliers [9].

Ali Shahbazi et al. [2] proposed a new tree similarity function with respect to tree structured data, namely extended sub-tree (EST) which maps sub-trees rather than tree nodes by utilizing new mapping rules. Three important steps in mining movement based communities

1. Creating a data structure for storing and managing all the trajectory details of one single user
2. Finding similarity measures between users
3. Finding movement based communities

Some of the most important user trajectory based applications are - friend recommendation, trajectory ranking and community based traffic sharing services and so on. The extensive application of tree structured data in today's information technology is obvious and trees can model many informative systems like XML and HTML and in many applications involving tree structured data, tree comparison is required [2]. An interesting problem on large trajectory data is the similarity search and a trajectory is represented as a sequence of locations each being associated with a corresponding time stamp [4]. User movement based communities are created by selection and using similarity measuring techniques such as

1. Longest Common Subsequence(LCSS)
2. Euclidian Distance
3. DTW
4. ERP
5. EDR

Trajectory classification has many useful applications such as city and transportation planning; road construction, design and maintenance; traffic congestion recognition; law enforcement; and home land security [13].

Most important methods that are used for community based clustering are – Clique method, Hierarchical method, and Betweenness method[22].Queries on historical trajectories may be offline or online; in offline processing it is generally permissible to have a relatively expensive preprocessing step on the set of trajectories to optimize query performance; online processing assumes that location updates arrive as a real-time data stream so that one must process queries in as the data is being updated [15].b)trajectory data indexing methods are classified into native space indexing methods and parametric space indexing methods [15].Trajectory queries retrieve nearby objects for a given route and such queries are useful in various domains including transportation and facility management [24].Trajectory queries have to handle

imprecise objects whose locations cannot be precisely determined and determining the query results over imprecise objects is technically challenging [24].

1.3 Applications of trajectory data mining

1) Air traffic controlling
2) Vehicle controlling
3) Finding the location of a particular pervasive device or vehicle
4) Finding the shortest route dynamically from a specific location

Recently, trajectory data mining has attracted much attention due to its wide applications. Trajectory data clustering and classifications are two important trajectory data mining techniques. Hnin Su Khaing et al. [8] proposed an efficient clustering algorithm which can solve the problem of DBSCAN for moving object trajectories and this algorithm consists of three phases – portioning, clustering and grouping. We define a trajectory stream as an totally ordered infinite sequences S={s1, s2, …, st, …} where si+1 comes after si and both si+1 and si are real values.

A key issue in the context of similarity search on uncertain trajectories is an appropriate distance function to measure the similarity between two uncertain trajectories [4]. Chunyang Ma et al. [4] proposed a p-distance measure which is a novel and adaptive metric to measure the dissimilarity between two uncertain trajectories. Uncertainty management is a central issue in trajectory databases and a common principle called location uncertainty is captured by a certain range centered on the position recorded in the database [10]. Hoyoung Jeung et al. [10] proposed a new model for capturing and representing the uncertainty of trajectory, termed evolving-density trajectory. Mobile devices are the potential sources of data providers for trajectory data management. A fundamental ingredient of trajectory data analysis tasks is the distance/similarity measure that can effectively determine the similarity of trajectories [8]. Popular methods that are used for finding distance/similarity measures are – Euclidian distance (ED), Dynamic Time Warping (DTW), Longest Common Sub Sequence (LCSS), Edit Distance with Real penalty (ERP), Edit Distance on Real Sequence (EDR) [8].

In modern era, automation is a key element. Automatically analyzing trajectory data records is a key element to assess the performances and the accuracy of new concepts of operations.

2. Related Work:

The main goal of the trajectory pattern data mining is to find hot regions from the trajectories, and then find sequential relationships among the hot regions, and then use these results in many real time applications such as clustering, classification, association and many other applications. Some authors have used association rules for storing movement behaviors of users and at the same time these association rules are represented in the trajectory pattern tree. Signature tree is also used by some authors as an indexing structure for deriving and managing sequential pattern relationship. Present study explains movement behaviors of users and then how these movement behavior details of users are used for clustering or classifying the users into groups.Examples for moving objects are - people vehicles, animals, and hurricanes etc. Many trajectory data mining based applications are beneficial to the universities, telecommunication industries, government, common people, business people and many commercial organizations. Comparing trajectories of two different objects is the fundamental function in trajectory data mining and as such it is very difficult to find a similarity metric for comparing trajectory sets of two different users because trajectories of different objects are constructed with different sampling strategies and sampling rates. Important layers in the trajectory data mining framework are - data collection, trajectory data mining techniques, applications [19].

Trajectory clustering techniques aim to find groups of moving object trajectories that are close to each other and have similar geometric shapes [16].Many works have elaborated on discovering user communities from user location history and such a community is thus referred to as a movement-based community [22].Some applications of movement based community are: Trajectory ranking, Community based traffic sharing services, and Friend recommendation; to discover movement-based communities, one should first formulate the similarity of users in terms of their trajectories [22].Wen-Yuan Zhu et al. [22] proposed a new similarity function to model user similarity in terms of their trajectory profiles as tree structures. Jing Yuan et al. [14] proposed a cloud-based cyber physical system for computing practically fast routes for a particular user using a large number of GPS-equipped taxis and the user's GPS-enabled phone. Jing Yuan et al. [14] proposed a landmark framework for the top-k road segments that are frequently traversed by taxi drivers according to the trajectory achieve. Jinfeng Ni et al. [15] proposed a new PA-tree indexing scheme for historical trajectory data, based on polynomial approximation, and showed how to apply PA-trees to support both offline and online processing of historical trajectories. Kai Zheng et al.

[16] proposed a framework for discovering all closed gatherings from a trajectory database. Xiaohui Li et al. [23] proposed a Group Discovery Framework for trajectory clustering. Zhiwei Lin et al. [26] tried to convert the tree similarity problem into a sequence similarity problem and it is simple and easy to measure sequential similarity measure than the tree similarity measure and there exist many sequence similarity measures than tree similarity measures. NingNan Zhou et al. [18] proposed a general Multi-Context Trajectory Embedding Model (MC-TEM), which provides a flexible way to characterize and leverage multiple contexts. Authors have proposed distributed representation method for modeling trajectory data by using deep learning and neural networks techniques. YinLai Jiang et al. [25] proposed a framework to extract embodied knowledge of human body motion by using Singular Value Decomposition (SVD). Nicholas Jing Yuan et al. [17] introduced the concept of latent activity trajectory, and generalized the problem of identifying functional zones using both location and mobility semantics, mined from latent activity trajectories. Potential applications of data mining are retailing, banking, credit card management, insurance, telecommunications, telemarketing and human resource management. A suitable distance function is used for finding similarity between uncertain trajectories. Trajectory clustering algorithms for moving objects are very useful in finding traffic jams, important location identification, and facilities available at a particular location at particular time etc. Some of the applications of trajectory data mining are Route Recommendation, Animal Migration, Transportation Management, Real time Traffic Information Details of Transport Organization, and Tourism. Some of the useful trajectory data mining applications are path discovery, shortest path discovery, individual behavior prediction, group behavior prediction, location prediction, service prediction and so on.

Path Discovery: Path discovery is also known as route discovery. There exist many ways for finding a path between any two given nodes. Sometimes there is a need to find most frequent path between two locations in a certain time period. For some applications, shortest path is needed, some applications require shortest path, another set of applications need cheapest or most popular path. Path discovery must find at least one path. Most frequent paths are better than the fastest paths or shortest paths in many real time applications .In terms of public transportation, people's real demand for public transportation are employed to identify and optimize existing flawed bus routes, thus improving utilization efficiency of public transportation. Dai et al. proposed a recommendation system that chooses different routes for drivers with different driving preferences. This kind of personalized route recommendation avoids flaws of previous unique recommendation and improves quality of user satisfaction. Previous experiences showed that human mobility as extraordinary regular and thus predictable [19].

One way is to detect common patterns and the other way is to detect outliers. Longest common subsequence (LCSS) clustering method uses similarity measure to cluster trajectory data patterns of objects. Longest common subsequence based clustering method appears to be more efficient and effective than Euclidean distance based measures and dynamic time warping distance functions, especially in the presence of noise.

For example, in a traffic management system, traffic jams may be determined by mining movement patterns of groups of vehicles (for example, groups of cars, buses, trains etc).

With respect to trajectory data management point of view, these applications require (1) a more structured recording of movements, which allows managing trajectories as efficiently as possible, and (2) different methods used to model and organize a set of trajectories. From a trajectory data analysis point of view, these applications ask for (3) trajectory real time analysis.

In the literature several popular methods have been proposed to improve trajectory data analysis, including the use of spatiotemporal databases and data mining techniques. These techniques can be classified according to the following groups:

1) Generation of trajectory patterns according to the geometrical properties of trajectory
2) Extraction of clusters of sample locations from trajectories, basically using time and space to determine trajectories located in dense regions, trajectories with similar shapes or distances, and trajectories that move between regions during the same time interval
3) Analysis of trajectories from a semantic point of view, trying to add context information
4) Development of architectures, ETL (Extraction, Transformation and Loading) processes, data models and languages for helping in the construction of trajectory data warehouses (TDWs).

Nowadays, many devices record traveling routes of users and vehicles as sequences of GPS locations.

A trajectory of a moving object is a set of different locations at various time instances. For, example, a trajectory, T, of a person (Rama) is represented as T = {(x1, y1, t1), (x2, y2, t2), …(xn, yn) where ti< ti+1. In general, two objects moving in the same path may or may not have the same trajectories.

Usually, the performance of the sequential data mining process is very much closely linked or associated with the length of sequence and the number of different locations appearing in the sequences. There is a need for representing and processing trajectories in a convenient, efficient and an effective optimal manner.

Advantages of trajectory pattern mining

1) Very easy to find traffic estimation on personalized or selected routes
2) Personalized graph construction
3) Finding the best road network
4) Finding the shortest route
5) Predicting destinations

3. Related Algorithm:

Wen-Yuan Zhu et al. [22] proposed a framework to mine movement based communities. They devised the framework into three phases that include i) constructing trajectory profiles of users ii) deriving similarity between trajectory profiles and iii) discovering movement based communities. Sequential probability (SP) tree as a data structure is used for representing user profile. Two algorithms named BF (breadth-first) and DF (depth-first) are used to construct SP-tree structures as user profiles. A Greedy algorithm named Geo-cluster is used to derive user communities.

Input: 1) A set of sequential probability trees {SP-tree$_1$, SP-tree$_2$, SP-tree$_3$,……, SP-tree$_n$}

2) Minimum similarity bound, δ

Output: A set of clusters representing user communities.

1) Construct a new connection graph G = (V, E) by using SP-tree$_1$, SP-tree$_2$, SP-tree$_3$,……, SP-tree$_n$ and δ
2) Initially consider all nodes of graph as separate clusters.
3) Previouscost= Total cost of cluster formation
4) Test=True
5) While (Test=True) do
6) Test =False
7) Initialize two clusters X_i, X_j each to empty
8) Minimum cost= Previous cost
9) For each pair of cluster C_i, C_j in the set c do
10) Present cost= previous cost+| C_i * C_j| - 2* Intercost (C_i, C_j)
11) If (present cost ≤ minimum cost) then
12) Test =True
13) Store C_i in X_i and C_j in X_j
14) End if
15) End for
16) If (Test=True) then
17) Combine X_i and X_j into a single cluster
18) Previous cost = minimum cost
19) End if
20) End while

4. Proposed Algorithms:

Algorithm 1: **Proposed** *Algorit*hm **for Clustering Trajectories of user movements**

Input: 1) A set of sequential probability trees {SP-tree$_1$, SP-tree$_2$, SP-tree$_3$,……, SP-tree$_n$}

2) Minimum similarity bound, δ

Output: A set of clusters representing user communities.

1) Construct a new connection graph G = (V, E) by using SP-tree$_1$, SP-tree$_2$, SP-tree$_3$,……, SP-tree$_n$ and δ
2) Initially consider all nodes of graph as separate clusters.
3) Previouscost= Total cost of cluster formation
4) Test=True
5) While (Test=True) do
6) Test =False
7) Initialize two clusters X_i, X_j each to empty
8) Minimum cost= Previous cost
9) For each pair of cluster C_i, C_j in the adjacent list of the profile graph G do
10) Present cost= previous cost+| C_i * C_j| - 2* Intercost (C_i, C_j)
11) If (present cost ≤ minimum cost) then
12) Test =True
13) Store C_i in X_i and C_j in X_j
14) End if
15) End for
16) If (Test=True) then
17) Combine X_i and X_j into a single cluster
18) Previous cost = minimum cost
19) End if
20) End while

A variant of above said algorithm, where a new similarity measure is considered is as follows:

Algorithm2:

Input: A set of 'n' number of customers' data in the form of trees, each tree represents a root consisting of a set of trajectories of user movements by a single user in the movement scenario.

Output: A set of clustered trajectories of user groups.

1. Create initially 'n' number of individual clusters such that each cluster consisting of a set of trajectories of items of one customer in the market basket.
2. Store full details of all the initial 'n' clusters of 'n' customers in the appropriate data structures.
3. For each individual cluster number i = 1 to n in terms of 1 do

4. For each individual cluster number j = i + 1 to n in terms of 1 do
   4.1 Find similarity measure between cluster i and j and store it in the appropriate efficient data structure for further processing.
   End-of-j for loop
5. Convert all computed similarity measures into normalized measures for ease and uniform processing.
6. Select the one with highest similarity measure value for clustering the two corresponding previous clusters.
7. End-of-i for loop
8. Repeat steps 3 through 7 until specified number of final clusters of trajectories of user behavior is Formed.

A trajectory data pattern represents the frequent movement behavior, which consists of:
1) Hot regions: areas with trajectories densely passed by
2) Relation: sequential relationships among hot regions

In Trajectory Data Pattern Mining, input is a set of trajectories and the output is a set of trajectory data patterns. Trajectory data patterns contain - Hot regions (areas an object usually stays) and relations (sequential or temporal relationships among hot regions).One can apply trajectory data mining techniques to mine user communities who have similar moving behaviors and then utilize these user communities in areas such as Geo-web sites for obtaining useful information such as finding interesting traveling/training paths, finding traveling/training partners and finding group buying behaviors.

Basic steps in trajectory data mining are illustrated below:

1) Preprocessing of trajectories
2) Hot regions are generated in this step
3) Find relationship between hot regions
4) How to organize patterns

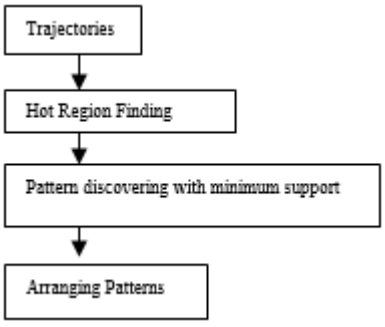


Fig. 1 Trajectory Data Mining Steps

Applications of Trajectory Data Patterns
1) Pattern aware trip planning:
2) Smart navigation:
   Community based traffic sharing platforms
   - Predicting your destination
   - Deriving personalized routes
   - Estimating traffic status
3) Location based Communities

Steps in discovering communities are:
1) Preprocess trajectories (finding hot regions),
2) Construct profiles of users (building trees such as sequential probability trees),
3) Formulate similarity measures, and then
4) Finally cluster users based on sequential probability

Original raw trajectories must be transformed and converted into valuable and suitable trajectories sequences. Trajectory patterns are useful to discover the movement behavior of users and then these movement behaviors are used to form movement groups of users. User movement behaviors are represented in terms of sequential relationships and probabilities. Creating user movement groups means clustering of trajectories of users.

5. Case study:

The trajectory data of taxy/ auto riksha movements in the Nellore town of Andhra Pradesh is considered for applying the proposed algorithm. The data is collected from about 425 taxi/ auto riksha drivers, considering twelve to fifteen journey trips of each driver. The user profiles are converted in to trees. These trees are termed as sequential probability trees, as they represent the sequence of user movements. Here the user is a taxi driver whose sequence of movements is recorded with respect to his pickup and travel sequence. From each individual user multiple journey records are noted where each record is a trajectory. he set of trajectories of a single user is used to design a user profile in the form of a tree.

Movement details of users are represented in the form of trajectory sets. For a given sets of trajectories of users, the main goal is to cluster trajectories of users into groups of trajectories of users. Trajectories in the same group have similar characteristics whereas trajectories in two different groups have dissimilar characteristics that are the most important goal is to cluster users into groups according to their own sets of trajectories.

Here 8 users are considered for explaining the computational procedure.

The user profile list is prepared based on individual list presented below:

User ID	Movement Trajectory
C101	ABCD,,BCD,ACD,AD
C102	MNOP,MOP,NOP
C103	MNOP,MOP,MP
C104	EFGH,EGH,IJKL,IKL
C105	UVWX,UWX,VWXY,VXY,RSTU,QST
C106	IKL,UVWX,UWX,VWX,QRSTU,RQS
C107	ABCD,ACD
C108	AD,EFGH,EGH,EH,FGH
C109	VWXY,VXY,QRTU,RTU
C110	QRSTU,QST,RSTU,RTU,QRTU,MP,RQS

For description of the methodology a profile graph for 10 users is considered as shown below:

The constructed profile graph is generated as a Geo-connection graph based on the objective function consisting of intra cost score and inter cost score. There exists geo connection relationship between two users if the similarity score between two profiles meet a threshold value that we assumed. The relationship is represented by drawing an edge between the corresponding nodes. It is considered that the nodes are processed in conventional increasing order of node counts in an efficient way for ease of processing and to reduce the computational time. For example, as in Fig (a), the cost of 1 to 7 is same as the cost of 7 to1. Hence1to 7 is only considered and 7 to 1 is the duplication of 1 to 7, hence it is not taken into consideration. Intra cost represents the minimum number of edges added to the nodes in the community (a clique which is the largest sub graph of a graph). Inter cost score is the minimum number of edges that must be removed from the existing geo-connection graph to convert all the communities as disjointed from each other. In Fig (a), all 10 nodes are distinct users. Its intra cost is 0 and inter cost is 11. Since 11 edges must be removed to make all the nodes as disjoint. In Fig (b) the intra cost is, since (5,6,9,10) is treated as one cluster, so if the six edges remove from the graph(b) then it becomes disjoint.

Cost between	(1,7)	(1,8)	(2,3)	(3,10)	(4,6)	(4,8)	(5,6)	(5,9)	(5,10)	(6,10)	**(9,10)**
Cost	10	10	10	10	10	10	10	10	10	10	**10**

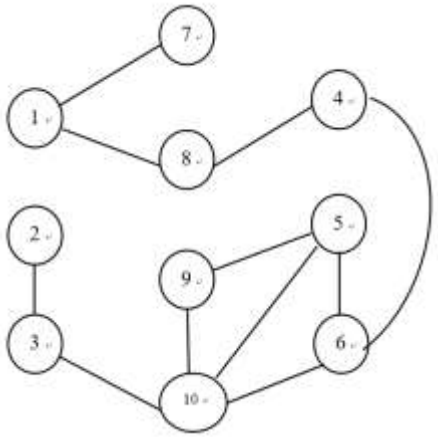


Fig. 2(a) Profile graph

Nodes	Nodes to be Traversed			Clustered Nodes	Intra Cost			Inter Cost		
1	7	8		1	1	1		1	1	
2	3			1	1			1	1	
3	10			1	1			1		
4	6	8		1	1	1		1	1	
5	6	9	10	1	1	1	1	1	1	1
6	10			1	1			1		
7				1						
8				1						
9	10			1	1			1		

6.Results :

Intermediate calculated values of the new-clustering algorithm are given as follows:

Table : Computation of Intermediate values of the New-Clustering Algorithm

Various costs are calculated as follows: Last cost =11;
Cost between (1,7) = current cost= last cost+1*1-2*1=11+1*1-2*1=11+1-2=10

Nodes	Nodes to be Traversed			Clustered Nodes	Intra Cost			Inter Cost		
1	7	8		1	1	1		1	1	
2	3			1	1			1		
3	10			1	2			1		
4	6	8		1	1	1		1	1	
5	6	9	10	1	1	2	2	1	1	1
6	10			1	2			1		
7				1						
8				1						
(9,10)	3	5	6	2	1	1	1	1	2	1

Cost between (1, 8) =11+1*1-2*1=11+1-2=12-2=10
(9,10) is minimum cost.
Now 9 and 10 nodes will be combined and now onwards 9 and 10 together treated as one unit with 2 nodes. That is (9,10) is a single component of the graph with 2 nodes. So, all the above tables are updated to reflect the fact that where ever 9 0r 10 is there, actually it is a single component containing two nodes.

In the next iteration the updated values are shown as follows :

Last cost = 10

Curr cost =last cost+|Ci*Cj|-2*costinter(Ci*Cj)

Cost of (1,7)=current cost= 10+1*1-2*1=10+1-2=11-2=9

Cost between	(1,7)	(1,8)	(2,3)	(3,10)	(4,6)	(4,8)	(5,6)
Cost	9	9	9	10	9	9	9
Cost between	(5,9)	(5,10)	(6,10)	⟨(9,10),3)	**((9,10),5)**	((9,10),6)	
Cost	10	10	10	10	**8**	10	

Nodes	Nodes to be Traversed			Clustered Nodes	Intra Cost			Inter Cost		
1	7	8		1	1	1		1	1	
2	3			1	1			1		
3	10			1	3			1		
4	6	8		1	1	1		1	1	
6	10			1	3			1		
7				1						
8				1						
(5,9,10)	3	6		3	1	1		1	2	

((9,10),5) is minimum cost.

Smallest cost is 8 so (9,10) and 5 are concatenated and then treated it as a single component and tabular values are updated as follows:

Last cost = 8

Curr cost =last cost+|C_i*C_j|-2*costinter(C_i*C_j)

Cost of (1,7)=curr cost=8+1*1-2*1=8+1-2=9-2=7

Cost of (1,8)= curr cost=8+1*1-2*1=8+1-2=9-2=7

Nodes	Nodes to be Traversed			Clustered Nodes	Intra Cost			Inter Cost		
1	7	8		1	1	1		1	1	
2	3			1	1			1		
3	10			1	4			1		
4	6	8		1	4	1		1	1	
7				1						
8				1						
(5,6,9,10)	3	4		4	1	1		1		

((5,9,10),6) is minimum cost.

Minimum cost is 7 so (5,9,10) and 6 are combined.

Cost of (1,7) = 7 + 1 * 1 – 2 * 1 = 7 + 1 -2 = 6

(4,8) is minimum cost.4 and 8 are combined.

Cost between	(1,7)	(1,8)	(2,3)	(3,10)	(4,6)	**(4,8)**	((5,6,9,10),3)	⟨(5,6,9,10),4)
Cost	6	6	6	9	9	**6**	9	9

Nodes	Nodes to be Traversed			Clustered Nodes	Intra Cost			Inter Cost		
1	7	8		1	1	2		1	1	
2	3			1	1			1		
3	10			1	4			1		
(4,8)	1	6		2	1	4		1	1	
7				1						
(5,6,9,10)	3	4		4	1	2		1	1	

Last cost = 6

Cost of (1,7) = 6 + 1*1 – 2*1 = 6+1-2 = 5

Cost between	(1,7)	(1,8)	((2,3),10)	((4,8),1)	((4,8),6)	((5,6,9,10),3)	((5,6,9,10),4)
Cost	4	5	11	5	11	11	11

(2,3) is minimum

Nodes	Nodes to be Traversed			Clustered Nodes	Intra Cost			Inter Cost		
1	7	8		1	1	2		1	1	
(2,3)	10			2	4			1		
(4,8)	1	6		2	1	4		1	1	
7				1						
(5,6,9,10)	3	4		4	2	2		1	1	

Last cost = 5

Cost of (1,7) = 5 + 1*1 – 2*1 = 5+1-2 =4

Cost between	(1,7)	(1,8)	((2,3),10)	((4,8),1)	((4,8),6)	((5,6,9,10),3)	((5,6,9,10),4)
Cost	4	5	11	5	11	11	11

(1,7) is minimum

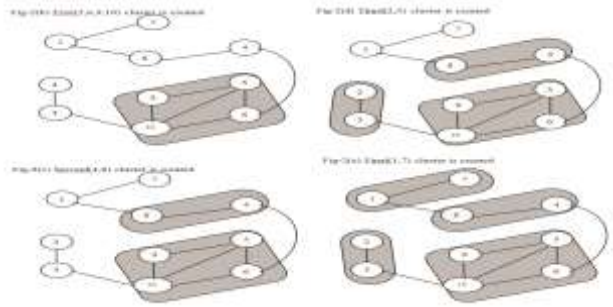
Last cost = 4

Cost of ((1,7),8) = 4 + 2*2 – 2*1 = 4+4-2 =6

Cost between	(1,7)	(1,8)	(2,3)	(3,10)	(4,6)	(4,8)	(6,10)	((5,9,10),3)	**((5,9,10),6)**
Cost	7	7	7	9	7	7	9	9	**7**

Cost between	((1,7),8)	((2,3),10)	((4,8),1)	((4,8),6)	((5,6,9,10),3)	((5,6,9,10),4)
Cost	6	10	6	10	10	10

(4,8),1) is minimum cost. But this minimum cost is greater than the previous minimum cost. Hence, the algorithm terminates. So, no further processing is required. **Final clusters for the proposed algorithm 1 are : (1,7), (2,3), (4,8), and (5,6,9,10)**



The same data is considered for applying a variant of the proposed algorithm2(one more proposal),in which different measures are used for finding similarity in which commonness among the clusters are measured using the formula:

Similarity(U_i , U_j) = intersection (U_i , U_j) / Union (U_i , U_j).

Where U_i , U_j are the set of user trajectories.

Similarity(1,2) = $U_1 \cap U_2$ /$U_1 \cup U_2$ = {ABCD,BCD,ACD,AD} ∩ {MNOP,MOP,NOP} / {ABCD,BCD,ACD,AD} ∪ {MNOP,MOP,NOP} = 0

Similarity(1,3) = $U_1 \cap U_3$ / U_1 ∪ U_3 = {ABCD,BCD,ACD,AD} ∩ {MNOP,MOP,MP}/{ABCD,BCD,ACD,AD}∪ {MNOP,MOP,MP} = 0

Similarity between (1,4)=0,(1,5)=0, (1,6)=0, **(1,7)=0.5**, (1,8)= 0.25, (1,9)=0, (1,10)=0, **(2,3)= 0.5**, (2,4)=0, (2,5)=0, (2,6)=0, (2,7)=0, (2,8)=0, (2,9)=0 ,(2,10)=0, (3,4)=0, (3,5)=0, (3,6)=0, (3,7)=0, (3,8)=0, (3,9)=0 ,(3,10)=0, (4,5)=0, (4,6)=1/9=0.11, (4,7)=0, (4,8)=2/7=0.28, (4,9)=0 ,(4,10)=0, (5,6)=2/10=0.2, (5,7)=0, (5,8)=2/7=0.28, (5,9)=2/8=025, (5,10)= 2/11 = 0.18, (6,7)=0, (6,8)=0, (6,9)=0, (6,10)= 1/12 = 0.08, (7,8)=0, (7,9)=0, (7,10)= 0, (8,9)=0, (8,10)= 0, (9,10)=2/9=0.22

Here (1,7) and (2,3) have highest similarity, hence these two are considered for clustering.

The same computations are performed between other pairs of interest. Final clusters for the proposed algorithm 2 are as shown below.

Cluster Number	Cluster Nodes	Movement Trajectory
1	1,4,7,8	ABCD,ACD,AD,BCD,EFGH,EGH,EH,FGH,IJKL ,IKL
2	2,3	MNOP,MOP,MP,NOP
3	5,6,9,10	IKL,MP,QRSTU,QRTU,QST,RQS,RSTU,RTU,U VWX,UWX,VWX,VWXY,VX

The final clusters for the said proposed variant are :{ (1,4,7,8), (2,3), (5,6,9,10) }

Results of existing process: The existing algorithm is also producing the comparable clustering out come as the result obtained from in proposed algorithm1. But the proposed algorithm is saving a significant amount of computation time, as the proposed algorithm is only considering adjacent pairs of the profile graph instead of all the pairs available in the profile set.

7. Comparisons:

The Geo-cluster algorithm makes use of each pair of objects from communities of users in the process of merging the objects. This requires C (n, 2) number of pairs to be compared where n is the number of initial independent user objects. To reduce this much of computation with lossless merging process we propose a novel framework in which, the pair with adjacent connectivity are only considered for merging process. This simple change will save reasonable amount of computation with no loss in the merging process.

Geo-Cluster algorithms compute two types of scores for community clustering. First score is called intra cost and the second score is called inter cost between two clusters C_i and C_j.

$$Cost_{total}(C) = \sum_{C_i \in C} Cost_{intra}(C_i) + \sum_{C_i, C_j \in C} Cost_{inter}(C_i, C_j) \qquad (1)$$

let us suppose that total number of nodes (users) is n and let a set of communities C be C_1, C_2,……, C_n

Another algorithm is also proposed to cluster the objects in which a better similarity measure is used.

O(n)	Number of comparisons for merging(Existing)	Number of comparisons for merging(Proposed)
5	10	8
10	45	22
15	105	36
20	190	52
25	300	69
30	435	115
35	595	175
40	780	260

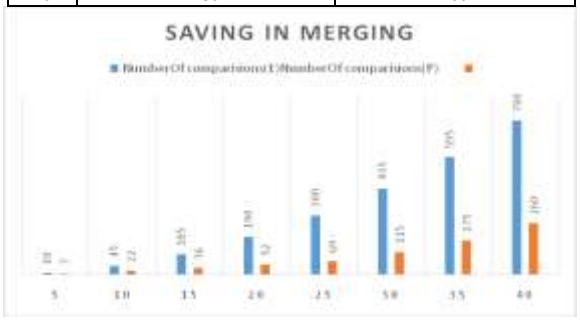


Fig. 4 Saving in Merging

The main numeric advantage of the proposed algorithm over existing algorithm is in terms of the number of comparisons used in the merging of the clusters. From the above graph it is clearly apparent.

O(n)	Run Time(E)	Run Time(P)
5	1.25	1.12
10	9.8	9
15	32.5	29.6
20	77	49
25	154.3	59
30	258	71
35	445	98
40	596	112.3

Fig. 5 Execution time

The above graph showed a significant amount of time difference between the two methods. The proposed algorithm is very economic in terms of the computation time.

8. Conclusion:

A new way of trajectory data clustering is opened here. This novel method is an improvement in terms of computational complexity reduction. To reduce this computational complexity with lossless merging process of clusters, we propose a novel framework in which, the pair with adjacent connectivity are only considered for merging process. This minute change saved reasonable amount of computation with no loss in the merging process and the final clusters formation. The novel process can be applied on the user profile trajectories of their movements to elicit the travel behavior. The trajectory data of taxy/ auto riksha movements in the Nellore town of Andhra Pradesh is collected and processed for evaluating the proposed algorithm. For users like taxi drivers or auto drivers covering a significant geographical area, the results of the clustering process provide them the most frequent and potential areas of their travel span. This facilitates the users to plan for better profitable regions of journey and improved Customer Relationship Management (CRM).

9. References :

[1] Aiden Nibali and Zhen He, "Trajic: An Effective Compression System for Trajectory Data", IEEE Trans On Knowledge and Data Engineering, Vol.27, No.11, November 2015, pages 3138-3151.

[2] Ali Shahbazi, and James Miller, "Extended Subtree: A new Similarity Function for Tree Structured Data," IEEE Trans. On Knowledge and Data Engineering, Vol.26, No.4, April 2014, pages 864-877.

[3] Banupriya C S, Dr.V.Vijaya Chamundeeswari, International Journal Of Engineering and Computer Science ISSN: 2319-7242 Volume 5 Issue 12 Dec. 2016, Page No. 19532-19635.

[4] Chunyang Ma, Hua Lu, Lidan Shou, and Gang Chen, " KSQ: Top-k Similarity Query on Unceratin Trajectories", IEEE Trans On Knowledge and Data Engineering, Vol.25, No.9, September 2013, pages 2049-2062.

[5] Costas Panagiotakis, Nikos Pelekis, loannis Kopanaki, Emmanuel Ramasso, and Yannis Theodoridis, "Segmentation and Sampling of Moving Object Trajectories Based on Representativeness", IEEE Trans On Knowledge and Data Engineering, Vol.24, No.7, July 2012, pages 1328-1343.

[6] C. Panagiotakis, N. Pelekis, and I.kopanakis, "Trajectory Voting and Classification Based on Spatio-Temporal Similarity in Moving Object Databases," Proc. Int'l Symp. Intelligent Data Analysis (IDA), pp.131-142, 2009.

[7] Gook-Pil Roh, Jong-Won Roh, Seung-Won Hwang, and Byoung-Kee Yi, "Supporting Pattern, Matching Queries over Trajectories on Road Networks," IEEE Trans On Knowledge and Data Engineering, Vol.23, No.11, November 2011, pages 1753-1758.

[8] Haozhou Wang, Han Su, Kai Zheng, Shazia Sadiq, and Xiaofang Zhou, "An Effective Study on Trajectory Similarity Measures," proceedings of the Twenty-Fourth Australian Database Conference (ADC 2013), Adelaide, Australia.

[9] Hnin Su Khaing, and Thandar Thein, "An Efficient Clustering Algorithm for Moving Object Trajectories", 3rd International Conference on Computational techniques and Artificial Intelligence (ICCTAI 2014) February 11-12, 2014 Singapore.

[10] Hoyoung Jeung, Saket Sathe, and Man Lung Yiu, "Managing Evolving Uncertainty in Trajectory Databases," IEEE Trans. On Knowledge and Data Engineering, Vol.26, No.7, July 2014, pages 1692-1705.

[11] Huey-Ru Wu, Mi-Yen Yeh, and Ming-Syan Chen, "Profiling Moving Objects by Dividing and Clustering Trajectories Spatiotemporally," IEEE Trans. On Knowledge and Data Engineering, Vol.25, No.11, November 2013, pages 2615-2628.

[12] Jae-Gil Lee, Jiawei Han and XiaoleiLi, "A Unifying Framework of Mining Trajectory Patterns of Various Temporal Tightness", IEEE Trans On Knowledge and Data Engineering, Vol.27, No.6, June 2015, pages 1478-1490.

[13] Jae-Gil Lee, Jiawei Han and Xiaolei Li, and Hong Cheng, "Mining Discriminative Patterns for Classifying Trajectories on Road Networks," IEEE Trans On Knowledge and Data Engineering, Vol.23, No.5, May 2011, pages 713-726.

[14] Jing Yuan, Yu Zheng, Xing Xie, and GuangZhong Sun, "T-Drive: Enhancing Driving Directions with Taxi Drivers' Intelligenc," IEEE Trans. On Knowledge and Data Engineering, Vol.25, No.1, January 2013, pages 220-232.

[15] Jinfeng Ni, and V.Ravishankar, "Indexing Spatio-Temporal Trajectories with Efficient Polynomial Appoximations," IEEE Trans On Knowledge and Data Engineering, Vol.19, No.5, May 2007, pages 1-16.

[16] Kai Zheng, Yu Zheng, Nicholas J. Yuan, Shuo Shang, and Xiaofang Zhou "Online Discovery of Gathering patterns over Trajectories", IEEE Trans On Knowledge and Data Engineering, Vol.26,No.8,August 2014, pages 1974-1988

[17] Nicholas Jing Yuan, Yu Zeng, Xing Xie, Yingzi Wang, and Hui Xiong, "Discovering Urban Functional Zones Using Latent Activity Trajectories", IEEE Trans On Knowledge and Data Engineering, Vol.27, No.3, March 2015, pages 712-725.

[18] NingNan Zhou, Wayne Xin Zhao, Xiao Zhang, Ji-Rong Wen, and Shan Wang, "A General Multi-Context Embedding Model for Mining Human Trajectory Data", IEEE Trans On Knowledge and Data Engineering, Vol.28, No.8, November 2016, pages 1945-1958.

[19] Tanuja.V MCA, M. Tech and Prof. P.Govindarajulu, M.Tech, Ph.D, S.V. University, Tirupati, "A Survey on rajectory Data Mining", International Journal of Computer Science and Security, Vol.10, Issue 5, 2016

[20] Tanuja .V MCA, M.Tech and Prof. P.Govindarajulu, M.Tech, Ph.D, S.V. University, Tirupati "Application of Trajectory Data Mining Techniques in CRM using Movement Based Community Clustering ", International Journal of Computer Science and Network Security, Vol. 16 , N0.11, November'2016

[21] Tanuja .V MCA, M.Tech and Prof. P.Govindarajulu, M.Tech, Ph.D, S.V. University, Tirupati "Application of Trajectory Data Clustering in CRM : A Case Study

" International Journal of Computer Science and Network Security, Vol. 17 , N0.1, January' 2017

[22] Wen-Yuan Zhu, Wen-Chih Peng, Chih-Chieh Hung, Po-Ruey Lei, and Ling-Jyh Chen, "Exploring Sequential Probability Tree for Movement-Based Community Discovery," IEEE TRANS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 11, NOVEMBER 2014, PAGES 2717-2730

[23] Xiaohui Li, Vaida Ceikute, Christian S. Jensen, and Kian-Lee tan, "Effective Online Group Discovery in Trajectory Databases," IEEE Trans. On Knowledge and Data Engineering, Vol.25, No.12, December 2013, pages 2752-2766

[24] Xike Xie, Man L. Yiu, Reynold Cheng, and Hua Lu, "Scalable Evaluation of Trajectory Queries over Imprecise Location Data," IEEE Trans. On Knowledge and Data Engineering, Vol.26, No.8, August 2014, pages 2029-2044.

[25] YinLai Jiang, Isao Hayashi, and Shuoyu Wang, "Knowledge Acquisition Method Based on Singular value Decomposition for Human Motion Analysis," IEEE Trans On Knowledge and Data Engineering, Vol.26, No.12, December 2014, pages 3038-3050.

[26] Zhiwei Lin, Hui Wang, and Sally McClean, "A Multidimensional Sequence Approach to measuring Tree Similarity," IEEE Trans. On Knowledge and Data Engineering, Vol.24, No.2, February 2012, pages 197-208.

V. TANUJA received Master of Computer Applications degree from Sri Venkateswara University, Tirupati, AP and Master of Technology degree in Computer Science & Engineering from Acharya Nagarjuna University, Guntur, AP. She is a research scholar in the department of Computer Science, Sri Venkateswara University, Tirupati, AP, India. Her research focus is on Data Mining in Customer Relationship Management.

P. GOVINDARAJULU, Retd. Professor, Department of Computer Science, Sri Venkateswara University, Tirupathi, AP, India. He received his M. Tech., from IIT Madras (Chennai), Ph. D from IIT Bombay (Mumbai), His area of research are Databases, Data Mining, Image processing, Intelligent Systems and Software Engineering.