Application of Trajectory Data Mining Techniques in CRM using Movement Based Community Clustering

V. Tanuja **P. Govindarajulu**

Department of Computer Science, S.V. University , Tirupati, AP, INDIA

Abstract:
Proliferation of location acquisition network technologies heightens very large trajectory data sets generation. Many data mining techniques have been proposed for efficient processing, managing and mining trajectory data. A trajectory is a sequence of geographical locations associated with timestamps. Day by day, trajectory data mining applications are increasing rapidly. Applications of trajectory data mining are – movement behavior analysis of objects, people, vehicles, animals etc, and finding geographic locations, finding desired paths geographically and so on. Suffix-tree-like index data structure can be used for efficient management of trajectories. Customer Relationship Management (CRM) is vital in present competitive business scenario. CRM requires information about the customer community to elucidate the behavioral patterns of the customers. Grouping of customers according to business inputs plays an important role. The clustering technique of Data Mining is a useful tool to grouping. Present work proposed a trajectory based clustering technique to pick up the customer groups from the customer data. The data source will be the data of a public transportation organization.

Keywords:
Data mining, CRM, Trajectory data mining, Trajectory datasets, Trajectory clustering.

1. Introduction

State-of-the-art positioning technology services such as Global Positioning System (GPS), Global System for Mobile communication (GSM), smart phone sensors etc. are available for finding the locations of desired moving objects geographically. A moving object can be a person, an animal, a vehicle, a mobile device and a trajectory of an animal describes its trace generated by daily activities such as walking, sleeping, eating and running etc. A trajectory of a vehicle is recorded by a GPS device installed in the vehicle and generally reports locations of the vehicle during fixed time intervals; for example, every second or every minute.

Road network and paths in the road network are referred whenever vehicle trajectories are considered. Transportation managers are interested in finding patterns about traffic jams, crowded roots, high demand roads etc.

Customer Relationship Management (CRM) is a term that refers to do policies and know-how that companies use to manage and analyze customer interactions and data throughout the customer lifecycle, with the goal of improving business relationships with customers, assisting in customer retention and motivating sales growth. CRM systems are designed to collect information on customers across different channels which could include the company's website, telephone, live chat, direct mail, marketing materials and social media. CRM systems can also give customer-facing staff detailed information on customers' personal information, purchase history, buying preferences and concerns. CRM is the buzzword of the present business environment without which the business entity cannot be imagined. In today’s customer centric business situation CRM is a common component in every business strategy. CRM requires collection, storage and analysis of customer centric data to go for decision making [3].

Data mining typically involves the use of predictive modeling, forecasting and descriptive modeling techniques as its key element. Using these techniques, an organization can able to manage customer retention(maintain), used to select the right prospects on whom to select, profile and segment customers(by identifying good customers), set optimal pricing policies, and objectively measure and rank which suppliers are best suited for their needs [14].

Major Goals the organization needs to achieve today include [3]

1. Cross selling the products.
2. Differentiating Loyal and Disloyal Customers.
3. Target Marketing to focus on prospective customers.
4. Prevention of defaults, bad loans.
5. To increase customer retention.

The following are the application of data mining in the customer relationship management system [11]

1) Customer Classification Analysis.
2) Customer Gaining Analysis.
3) Customer Losing and Maintaining Analysis.
4) Customer Profit-making Ability Analysis and Forecast.
5) Cross Selling Analysis.
6) Customer Satisfaction Analysis.
7) Customer Credit Analysis.

Knowledge discovery applications are emerging in a variety of industries [4]

Customer segmentation—All industries can take advantage of data mining to discover discrete segments in their customer bases by considering additional variables beyond traditional analysis.

Manuscript received November 5, 2016
Manuscript revised November 20, 2016

Manufacturing—Through choice boards, manufacturers are beginning to customize products for customers; therefore they must be able to predict which features should be bundled to meet customer demand.

Warranties—Manufacturers need to predict the number of customers who will submit warranty claims and the average cost of those claims.

Frequent flier incentives—Airlines can identify groups of customers that can be given incentives to fly more.

In order to build good model of CRM system, there are a number of steps that need to be followed. Following are the basic steps of data mining for effective CRM[5] -

- Define Business Problem
- Build Marketing Database
- Explore Data
- Prepare Data For Modeling
- Build Model
- Evaluate Model
- Deploy Model and Results.

Data mining tools helps CRM by providing the framework, which covers: i) to analyze the business problem ii) to prepare the data requirements iii) to build the suitable model with respect to business problem and, iv) to validate and evaluate the designed model [10].CRM consists of the following dimensions :(1) Customer Identification; (2) Customer Attraction; (3) Customer Retention; (4) Customer Development. These four dimensions can be seen as a closed cycle of the CRM system.

Clustering is used to group a number of records into intra similar and inter dissimilar sections. This grouping is done by measuring the similarity among the objects of interest. Several methods have been evolving to do clustering. With respect to CRM clustering offers the grouping of customers of a particular business entity from which managerial decision making can be whitened.

Trajectory based information can be filtered to obtain similar groups using clustering. This knowledge can be used to identify vital groups to get a practical or business edge.

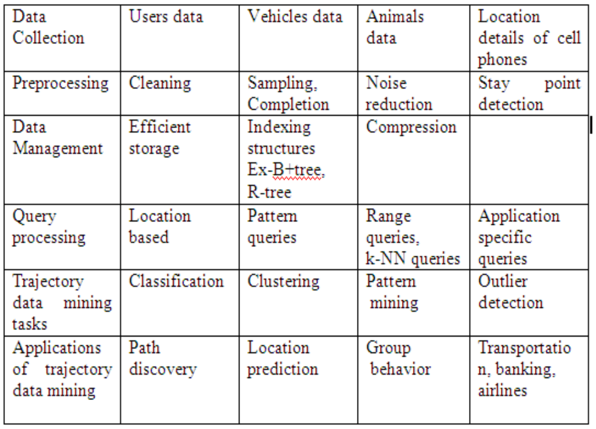
A trajectory pattern is a sequence of places of interests which a user frequently visits. Trajectory sequential patterns can be explored in many different ways. A trajectory is represented by a sequence of time stamped points. Data mining technology is useful for extracting important information and knowledge from trajectory data. Clustering of users' trajectories and giving appropriate meanings to user movements are active areas of research. Markov chain model is one of the best frame works for representing trajectory data.

Primary goal of present study is to find efficient and effective means and ways for analyzing user movement trajectories and then extracting useful user preferences from the corresponding trajectories. Many clustering algorithms exist for mining user GPS trajectories. The purpose of this paper is to propose a way to extract user preferences by clustering multiple users based on their trajectories. Clustering means to identify similar and dissimilar groups in a given dataset.

Along with the introduction, the rest of the paper is organized into five sections. In the second section the related work in trajectory data mining is explained together with the framework and the steps of the process in general. The third section reflects the methodology in detail. The proposed algorithm and experiments are followed in the remaining sections.

2. Related Work

The field of trajectory data mining is spreading itself with fast evolution in it. Here efforts are made to mention the developments in the field. The following table exploits the trajectory data mining paradigm.



Data Collection	Users data	Vehicles data	Animals data	Location details of cell phones
Preprocessing	Cleaning	Sampling, Completion	Noise reduction	Stay point detection
Data Management	Efficient storage	Indexing structures Ex-B+tree, R-tree	Compression	
Query processing	Location based	Pattern queries	Range queries, k-NN queries	Application specific queries
Trajectory data mining tasks	Classification	Clustering	Pattern mining	Outlier detection
Applications of trajectory data mining	Path discovery	Location prediction	Group behavior	Transportation, banking, airlines

Fig-1 A framework for clustering trajectory data mining

Location acquisition and mobile computing technologies are generating very large spatial trajectory data. Large spatial trajectory datasets represent mobility of many objects such as people, vehicles, and animals etc. Trajectory data mining is rapidly gaining its popularity in many spatial database applications. There exist a wide spectrum of applications driven and improved by trajectory data mining, such as path discovery, location or destination prediction, movement behavior analysis for individual or a group of moving objects, making sense of trajectories and other applications of urban service . A trajectory of a moving object is a discrete trace that the moving object travels in geographical space. Privacy-preserving is a crucial problem in every procedure of trajectory data mining [20]. Many methods exist for transforming trajectories into other formats such as trees, graphs, and matrices as these data structures are convenient for applying many data mining and machine learning techniques efficiently and effectively.

An index structure called TrajTree is developed to manage trajectory data especially for retrieval tasks like k-NN queries [20]. A location-based query attempts to find trajectories that are close to all query locations where the query is a small set of locations with or without a specific order constraint and one typical application is route recommendation for a trip to multiple places [20].In the present study trajectory relationships are related to vehicles and roads. Potential applications of data mining are –banking, insurance, credit card management, telecommunications, retailing, telemarketing and human resource management.

Trajectories of moving objects is a new kind of spatio-temporal data generated by mobile devices and a general framework has been proposed for modeling trajectory patterns during the conceptual design of a database [15].

The field of trajectory data mining is conveniently represented in Fig. 1 [18].

The huge volume of spatial trajectories enables opportunities for analyzing the mobility patterns of moving objects, which can be represented by an individual trajectory containing a certain pattern or a group of trajectories sharing similar patterns . Mobility of people: People have been recording their real-world movements in the form of spatial trajectories, passively and actively, for a long time [18].

Mobility of transportation vehicles: A large number of GPS-equipped vehicles (such as taxis, buses, vessels, and aircrafts) have appeared in our daily life. For instance, many taxis in major cities have been equipped with a GPS sensor, which enables them to report a time-stamped location with a certain frequency. Such reports formulate a large amount of spatial trajectories that can be used for resource allocation [18]. Mobility of animals: Biologists have been collecting the moving trajectories of animals like tigers and birds, for the purpose of studying animals' migratory traces, behavior, and living situations [18].Mobility of natural phenomena: Meteorologists, environmentalists, climatologists, and oceanographers are busy collecting the trajectories of some natural phenomena, such as hurricanes, tornados, and ocean currents.

Typically, trajectory data are obtained from mobile devices that capture the position of an object at specific time intervals [8].The motivation for mining trajectory datasets is the possibility of realizing inherent information, helping to gain understanding of the fundamental phenomena of movement. The understanding of movements is helpful within many contexts. Increasing availability of data and the number of methods for utilizing this data is making trajectory data mining gain sufficient importance in various domains, including urban planning, traffic flow control, public health, wildlife protection and location aware advertising [9].

The recent advances in technologies for mobile devices, like GPS and mobile phones, are generating large amounts of a new kind of data: trajectories of moving objects and these trajectory data can be used in a variety of applications. For example, trajectories obtained from GPS devices of car drivers can be used for traffic management, for urban planning, for insurance companies and so on . Trajectory data mining algorithms are mainly based on trajectory similarity [7]. There is currently a huge amount of data being collected about movement of objects. Such data is called spatiotemporal data and paths left by moving objects are called trajectories [2]. Recently researchers have been targeting those trajectories for extracting interesting and useful knowledge by means of pattern analysis and data mining [2]. One of the main problems for business today, however, is that such raw trajectory data leads to little knowledge for decision makers if it is given to the business decision makers without analyzing them and extracting useful knowledge [2]. A trajectory-mining application problem is an issue that can be solved by mining trajectories such that the solution is useful in application fields [6]. The data that represent the movement of an object have been referred to using several different terms including trace data or traces, movement data, mobility data, and trajectory data or trajectories [6]. Similar to the general domain of data mining, trajectory data mining aims at discovering interesting patterns from the data and it has two primary goals: prediction and description. Prediction consists in using some variables in the data to determine unknown or future values of other variables of interest, while description focuses on finding human-interpretable structures describing the data [6]. Given the advent of telecommunications and particularly cellular phones, GPS technology and satellite imagery, robotics, Web traffic monitoring, and computer vision applications, the amount of spatiotemporal data stored every day has grown exponentially [12].

A number of trajectory classification methods have been proposed mainly in the fields of pattern recognition, bioengineering and video surveillance [12]. Nowadays, wireless networks and GPS are the two important sources of trajectory data for moving objects. Telecommunication companies accrue masses of cell-based movement data. Also the technologies like GPS provide a considerably more precise positioning. Yet, the trade-off for high-quality data lies in substantially reduced quantity as GPS data are not easily available. Data mining on spatiotemporal data, trajectory data in particular, is a largely unexplored area [12].Zhixian Yan et al. [21] proposed a trajectory ontology framework to capture semantics for trajectory data as well as to support automatic reasoning. Zheng K., et al. [21] proposed an algorithm on representing the uncertainty of the objects moving along road networks as time dependent probability distribution functions.

Range queries are associated with specific lower and upper bounds of the range. Uncertain queries are usually modeled by using range queries and probability density functions (pdf). Nearest neighbor query is one of the most important queries in spatial-temporal trajectory data mining. Another type of query is finding k-most important similar trajectories (top-k) for a given trajectory data set. Pattern queries generally represent patterns in the query as regular expressions. Keyword queries for semantic trajectories are particularly useful for tourists in trip planning. Most important steps in trajectory data management are:

Trajectory data collection

Trajectory data are generated by various moving objects and collected from multiple data sources. Then, main part of trajectory mining techniques are presented with five components, i.e., preprocessing, data management, query processing, trajectory data mining tasks, and privacy protection. Finally, in the layer of applications, it is reviewed about an extensive set of applications from six categories [20].

Trajectory preprocessing

During trajectory preprocessing trajectory data are cleaned, modified, pruned, segmented, calibrated and sampled conveniently. Preprocessing improves the quality of the trajectory data [20].

Trajectory data management

Trajectory data are simplified before being stored in the memory. Efficient, effective and scalable storage systems must be provided for trajectory data storage. Appropriate fast and scalable index data structures are also necessary to support for fast query processing and interactive query processing. Storage of huge amount of trajectory data is the main problem in trajectory data mining. To manage trajectory data special type of indexing tree structures are needed [20] .

Trajectory data mining

There are four main categories of objects constituting the majority of trajectory data; 1) human, 2) transportation, 3) animal, and 4) natural phenomena. Human trajectory data is concerning the movements people do as they travel by foot. Transportation trajectory data is also connected to humans moving but is specifically concerning movements made with vehicles [19].

Application of mined knowledge

Authors divided the trajectory knowledge into main groups - knowledge resource, knowledge types, knowledge datasets, data mining tasks, data mining techniques and applications used in knowledge mining [8]

Query processing

Different types of queries have to be processed to retrieve desired data from desired locations geographically, e.g., location-based queries, range queries, nearest neighbor queries, top-*k* queries, pattern queries, aggregate queries and other application specific queries. These queries are processed based on an underlying storage system and index structure. A location-based query attempts to find trajectories that are close to all query locations where the query is a small set of locations with or without a specific order constraint. Range queries retrieve the trajectories falling into (or intersecting) a spatial (or spatiotemporal) range. One typical application is route recommendation for a trip to multiple places.

Trajectory data mining tasks: Trajectory data mining tasks are divided into the following categories:

Pattern Mining: Pattern mining is analyzing the mobility patterns of people, objects, animals, cell phones, and vehicles etc. Different types of patterns are - sequential patterns, group patterns and periodic patterns etc. Regarding each trajectory as a sequence, a sequential pattern is often defined as a subsequence that has at least k trajectories that share the subsequence, where k is a user specific threshold value. Periodic pattern explain the behavior of moving objects. Probabilistic models and reference locations play an important role in modeling periodic patterns. Trajectory patterns can be discovered from a single trajectory or a group of trajectories.

Clustering

Trajectory clustering is useful for dividing trajectories into groups with similar movement patterns. Mobility based clustering is essentially forming the similarity groups of moving vehicles, people, and cell phones etc. Many general frameworks exist for mining communities from multiple sources of trajectories. Moving objects are clustered based on trajectory related information such as semantic meaning of trajectories, movement velocity, feature movements, feature characteristics, temporal duration, and spatial dispersion. The mobility-based clustering is less sensitive than the density-based clustering to the size of trajectory dataset [20].

Classification

Trajectory classification means building a trajectory classifier model from the trajectory training data set and then using the classifier model to determine the class label of the new trajectory tuple.

Outlier detection

Trajectory outliers can be items that are significantly different from other items in terms of some similarity metric. It can also be events or observations (represented by a collection of trajectories) that do not conform to an expected pattern (e.g., traffic congestion caused by a car accident).

3. Methodology

Clustering of customers is useful in many business areas such as location based services and trajectory recommendation services. Movement details of customers are represented using trajectory profiles of customers.

Trajectory profiles of customers are constructed before clustering or classification of customers based on their movement behavior. Location aware sensors attached with global positioning system (GPS) devices are very much convenient for constructing trajectory profiles of customers and these trajectory profiles of customers are the basic building blocks of customer relationship management. The sequences of activities of customers are called customers' trajectories. Trajectories represent sequences of real life activities of customers. Websites that utilize these trajectories of customers are growing rapidly and becoming popular. Movement based community technique clusters customers based on customer activities or customer locations. Many applications that use sequences of customer activities are developing in many organizations.

Example applications based on sequences of customer activities are:

1. Friend recommendation.
2. Trajectory based ranking technique.
3. Community based traffic sharing services.

1. Friend recommendation.

Friendly relationships among the customers are created based on sequences of activities of customers. Customers are clustered using their trajectory profiles. Trajectories shared by customers are potential sources of friend recommendation.

2. Trajectory based ranking technique

Trajectory profiles of customers are stored in websites and most of the websites are already equipped with query interfaces for retrieving customer trajectories. In general customer trajectories are retrieved based on a keyword or a set of keywords. Trajectory ranking is very useful for clustering whenever query returns a very large size of dataset of trajectories. Trajectory ranking is also useful for pruning the customer trajectories returned by the query. Shared trajectory details of customers are particularly useful for clustering customers in many business applications.

3. Community based traffic sharing services.

Traffic sharing facility is one of the recently available services that are becoming popular and useful in many applications with the help of fast developing mobile applications. To discover activity based clusters of customers, one best way is first analyze all the shared customer trajectories to find similarity behaviors of customers and then apply clustering technique. Many similarity measuring functions are available to measure similarity degree of customers based on sharing details of activities of customer trajectories. Activities of trajectories are recorded using timestamps in chronological order. All most all trajectory data are time series data. Whenever trajectories are collected at very first time, they may contain data uncertainty and as well as noise and as such they are not useful directly for clustering of customers using directly available similarity measuring functions. Raw trajectories must be converted into useful trajectories by removing noise and uncertainty present in the originally collected raw trajectories. Processing cost of raw trajectories is very high. Only frequent sequences of activities of customers are considered to reduce computing cost of finding similarity measures. Converted trajectories are called transformed trajectories which contain only important sequence of activities of customers. An identical sub-tree represents a similar sub structure between trees, whereas disjoint mapped nodes indicate no similar structure between the two trees [1].

One way to determine characteristics of each group from a set of transformed trajectories is use sequential pattern mining methods to obtain frequent sequential patterns. Characteristics of sequence of activities are represented effectively using both sequence of activities and probabilities of corresponding sequences. Trajectory profile of each customer is represented by a special data structure called sequential activities probability tree (SAPT). Here, n numbers of SAPTs are required to represent trajectory profiles of n customers.

Main steps of the present work are:

1. Creating sequential activities probability trees
2. Finding similarities between trajectory profiles of customers
3. Clustering of customers based on similarity measures

1. Creating sequential activities probability trees (SAPTs)

One sequential activities probability tree will be constructed for one customer trajectory profiles. Trajectory profiles of n-customers are represented using n-SAPs. Trajectory trees capture not only the sequential movement activities of customers but also the transition probabilities among activities. The sequential patterns among the customer activities specify the frequency of occurrence of customer activities. SAPT mines both sequential patterns and transition probabilities. SAPT is an m-way nonlinear data structure and each node is labeled by tree edges traversed from the root node of the SAP tree, and labels represent sequential patterns. Each node is associated with a conditional table that reflects the next movement probabilities. Transmission probabilities are derived by traversing SAPT starting from the root node. Customer profiles are compactly stored in the SAPT. Sequential patterns of customer activities and the corresponding transition probabilities are effectively represented in SAP tree model whose time complexity of insertion, search and deletion operation is O(log n).

2. Finding similarities between trajectory profiles of customers

Trajectory profiles of customers are directly represented in SAP trees. First, n-number of SAPT trees are constructed for n-customer activities. Similarity measures between any two SAP trees are computed based on different types of similarity measuring techniques or scores. To find similarity between two SAP trees different details of the SAP trees must be considered. Possible details of SAP trees are:

- Total number of branches,
- Total number of nodes
- Total number of string matched branches
- Total number of not string matched branches
- Total number of common nodes
- Total number of distinct nodes
- Total number of branches
- Support values of tree nodes and
- The conditional tables of the tree nodes etc.

Also certain types of mathematical as well as statistical based similarity measuring techniques are also available for comparison.

The data for this work is a weighted trajectory. Here weight is nothing but a pre-defined threshold. A weighted trajectory is a sequence or subsequence of an object movement data which covers a threshold. APSRTC transportation data is considered .A bus route with a list of stations covered by a bus is considered. Here a route is a collection of stations in a sequence.

A subset of a route in a single day/week/month is a weighted trajectory if all of its constituent stations satisfied with a threshold number of passengers.

The following is the format of the data may be considered.

Route Id	Vehicle NO	Route sequence	Occupancy sequence
100	467	ACDEFH	45-56-23-44-25-25
100	468	ACDEFH	45-56-23-44-25-25
100	469	ACDEFH	45-56-23-44-25-25
100	469	ACDEFH	45-56-23-44-25-25
101	564	ADEFH	52-23-49-40-40
101	564	BFHKLST	41-22-23-48-48-40-40
101	565	ACDEFG	45-33-45-44-41-41
102	666	ACDEFH	45-56-23-44-26-26
102	667	ACDEFH	45-52-23-44-25-25
102	667	ACDEFHIJKLM	45-56-23-44-25-29-52-35-55-44-44

Suppose the good occupancy threshold is 40, then a weighted trajectory is

ACE, AEFH, BKLST, ADEFG, ACE, ACE, ACEIKLM

Such weighted trajectories are the input to the clustering process. The main components of the clustering process include the tree construction for each movement group (route), similarity search among groups, creating new groups (clustering).

The above data set after clustering forms the new groups of trajectories satisfying the threshold value. The new grouping is different from the basic group as some of the trees are merged to another tree increasing that may increase homogeneity of the clusters formed.

The weighted trajectory sets can be formed for various weights.

These groups represent the high demand routes. A public transportation company like APSRTC can use this group information to take decisions on new route planning which may attract the passengers towards the company services. A trajectory set with poor weights may also be considered to uncover weak performing routes. This information can be used to take decisions to discontinue certain routes or to attract customers with some offers to strengthen the route performance with respect to occupancy.

4. Algorithm

The present section covers two algorithms. One is the popular Breadth First Search method and the second one is the proposed algorithm. Breadth First Algorithm for constructing Sequential Activity Probability Tree[16]:

Algorithm 1: Breadth First Method (BFM)

Input:

1. Input is a set of profiles of n users, and T represents transformed trajectories.
2. Given minimum conditional probability threshold.
3. Given minimum support threshold.

Output: Only a single sequential activity probability tree (SAPT) that represents profile of a single user.

```
Root = null
S ={root}
K=0
While ($S_k \neq 0$) do
  $S_{k+1}$=0
  For each node value s in the set $S_k$ do
    Find frequent hot regions and then create conditional
    table of node s
      For each of ⅄ in frequent hot regions do
        If ⅄ is in the conditional table of node is s then
        Create a new trajectory set called sσ
        S⅄ is a child of s, so add node sσ into $S_{k+1}$
      End if
    End for
  End for
  k = k + 1
End of while
```

Proposed Algorithm for Clustering:

Present study proposes a new algorithm called New Clustering for clustering customers.

Algorithm New-Clustering

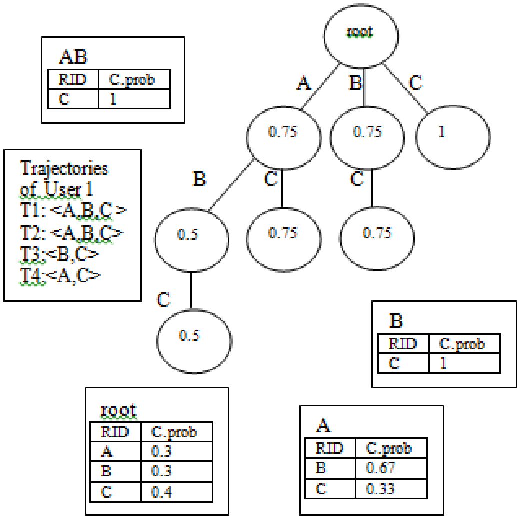
Input: A set of n number of sequential activity probability trees and a threshold values

Output: A set of clusters

```
1. for each cluster number  i = 1 to n do
2.  for each cluster number  j = i + 1 to n do
3. find similarity measures between clusters i and j and
   store in appropriate data structure for future processing
4.  end of for loop
5. Sort all the normalized measures and then select highest
   value for clustering the two corresponding previous
   clusters.
6. Repeat the steps 1 through 5 until threshold is satisfied
```

To explain the above algorithm, a hypothetical example is followed.

Trajectory(**SEQUENTIAL PATTERNS OF USERS)**



AB

RID	C.prob
C	1

B

RID	C.prob
C	1

root

RID	C.prob
A	0.3
B	0.3
C	0.4

A

RID	C.prob
B	0.67
C	0.33

Fig-1. User U1's trajectories and SAP-tree SAPT1

Nodes of SAPT$_1$ are named as shown below

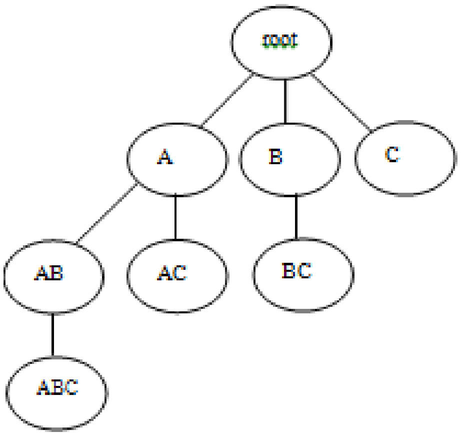


Fig-2 Node numbering convention

User 1's Trajectories are : T1: <A,B,C >, T2: <A,B,C>, T3:<B,C>, T4:<A,C>

Support computational details of User1's trajectories at the 1st level are :

Support of hot region is computed as,

$$support = \frac{No.of\ Trajectories\ of\ the\ hot\ region}{Total\ trajectories}$$

For Example, Support of node A =

$$\frac{No.of\ Trajectories\ in\ which\ A\ present}{Total\ trajectories} = \frac{3}{4} = 0.75$$

Similarly support details of B and C respectively are 3/4=0.75 and 4/4=1

Support count details at the second level are :

$$support\ of\ node\ AB = \frac{No.of\ Trajectories\ in\ which\ node\ AB\ present}{Total\ trajectories}$$

Support values for all nodes in all levels are computed in the similar fashion.

Conditional Probability (C. Prob) at root node are :
Totally there are 10 distinct hot regions. Out of 10, there are 3 A's, Therefore conditional probability of A=3/10=0.3. Similarly conditional probabilities of B and C respectively are 3/10, 4/10.

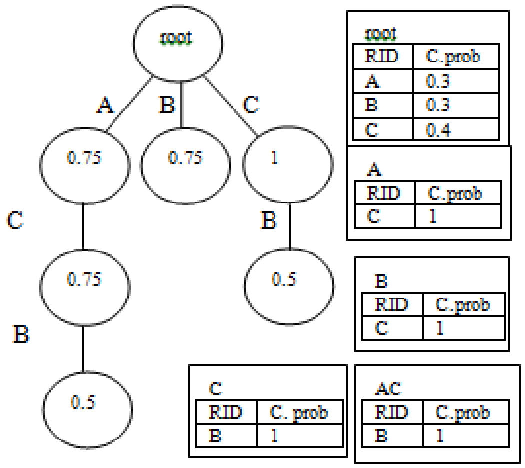
Conditional probability of node A to B =

$$\frac{No.of\ Trajectories\ starting\ with\ A\ and\ ending\ with\ B}{No.of\ Trajectories\ Starting\ with\ A} = 2/3=0.66$$

Conditional probability of node A to C =

$$\frac{No.of\ Trajectories\ starting\ with\ A\ and\ ending\ with\ C}{No.of\ Trajectories\ Starting\ with\ A} = 1/3=0.33$$

Similarly C.Prob for other nodes are computed.



root

RID	C.prob
A	0.3
B	0.3
C	0.4

A

RID	C.prob
C	1

B

RID	C.prob
C	1

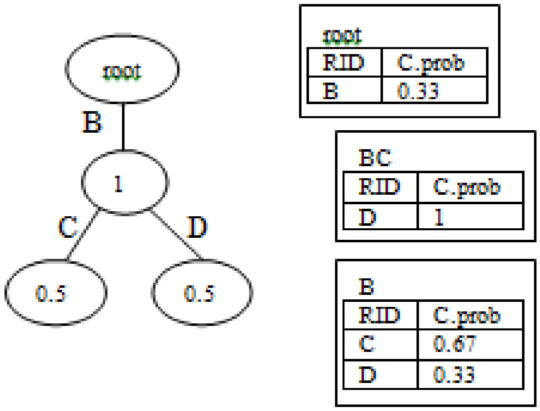
C

RID	C. prob
B	1

AC

RID	C.prob
B	1

Fig-3. User U2's trajectories and SAPT-tree SAPT2



root

RID	C.prob
B	0.33

BC

RID	C.prob
D	1

B

RID	C.prob
C	0.67
D	0.33

Fig-4. User U3's trajectories and SP-tree SAPT3

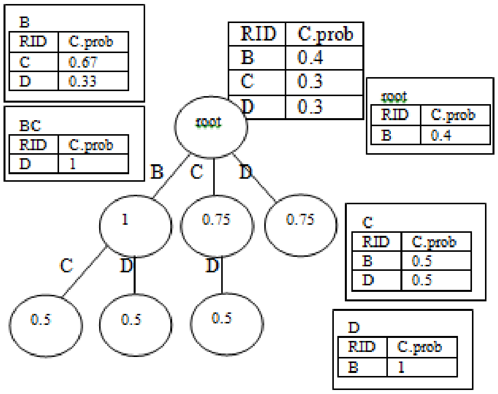
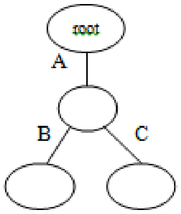


Fig-5. User U4's trajectories and SP-tree SAPT4



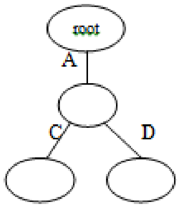


Fig-6 SAPT5Fig-7 SAPT6

Movement behaviors are characterized using movement sequential patterns and transition probabilities [16]. We propose a special and new similarity measure finding algorithm for grouping users into clusters. This new similarity uses a simple normalized measure for comparing two users. Based on the minimum threshold value clusters are created. Clustering process stops when all users are checked. Trajectory data management [17] requires efficient, effective, robust, accurate and scalable features for storing very large sizes of trajectory training data sets. Efficient storage of very large trajectory databases is the fundamental problem in trajectory management [17].

5. Data Source and Experiments

Trajectories that satisfy the specified threshold values are taken for constructing sequential activities probability trees and then these trees are given as input to the proposed clustering algorithm. The input consists of trees derived from trajectories of each sequential activity probability tree as mentioned in the following format.

Node names of $SAPT_1$ ={root, A,B, C, AB, AC, BC, ABC}
Node names of $SAPT_2$ = {root, A, B, C, AC, CB, ACB}
Node names of $SAPT_3$ = {root, B, BC, BD}
Node names of $SAPT_4$ = {root, B, C, D, BC, BD, CD}
Node names of $SAPT_5$ = {root, A, AB, AC}
Node names of $SAPT_6$ = {root, A, AC, AD}

The output includes clustered trees followed by final user groups as follows.

$SAPT_1 \cap SAPT_2$= {root, A, B, C, AC}
$SAPT_1 \cup SAPT_2$= {root, A, B, C, AB, AC, BC, ABC, CB, ACB}
Normalized similarity measure between trees $SAPT_1$ and $SAPT_2$ =$\frac{SAPT_1 \cap SAPT_2}{SAPT_1 \cup SAPT_2} = \frac{5}{10}$ = 0.5, $SAPT_1 \cap SAPT_3$= {root, B, BC}
$SAPT_1 \cup SAPT_3$= {root, A, B, C, AB, AC, BC, ABC, BD}
Normalized similarity measure between trees $SAPT_1$ and $SAPT_3$ =$\frac{SPT_1 \cap SPT_3}{SPT_1 \cup SPT_3} = \frac{3}{9}$ =0.33, $SAPT_1 \cap SAPT_4$={root, B, C, BC}
$SAPT_1 \cup SAPT_4$={root, A, B, C, AB, AC, BC, ABC, D, BD, CD}
Normalized similarity measure between trees $SAPT_1$ and $SAPT_4$ =$\frac{SAPT_1 \cap SAPT_4}{SAPT_1 \cup SAPT_4} = \frac{4}{11}$ =0.363,
$SAPT_1 \cap SAPT_5$= {root, A, AB, AC}
$SAPT_1 \cup SAPT_5$= {root, A, B, C, AB, AC, BC, ABC}
Normalized similarity measure between trees $SAPT_1$ and $SAPT_5$= $\frac{SPT_1 \cap SPT_5}{SPT_1 \cup SPT_5} = \frac{4}{8}$ =0.5, $SAPT_1 \cap SAPT_6$ ={root, A, AC}
$SAPT_1 \cup SAPT_6$= {root, A, B, C, AB, AC, BC, ABC, AD}
Normalized similarity measure between trees $SAPT_1$ and $SAPT_6$ =$\frac{SAPT_1 \cap SAPT_6}{SAPT_1 \cup SAPT_6} = \frac{3}{9}$ =0.33, $SPT_2 \cap SPT_3$= {root, B}
$SAPT_2 \cup SAPT_3$= {root, A, B, C, AC, BC, BD, CB, ACB}
Normalized similarity measure between trees $SAPT_2$ and $SAPT_3$= $\frac{SAPT_2 \cap SAPT_3}{SAPT_2 \cup SAPT_3} = \frac{2}{9}$=0.222, $SAPT_2 \cap SAPT_4$= {root, B, C}
$SAPT_2 \cup SAPT_4$= {root, A, B, C, AC, CB, ACB, D, BC, BD, CD}
Normalized similarity measure between trees $SAPT_2$ and $SAPT_4$ =$\frac{SAPT_2 \cap SAPT_4}{SAPT_2 \cup SAPT_4} = \frac{3}{11}$ =0.272, $SAPT_2 \cap SAPT_5$={root, A, AC}
$SAPT_2 \cup SAPT_5$= {root, A, B, C, AC, AB, CB, ACB }
Normalized similarity measure between trees $SAPT_2$ and $SAPT_4$= $\frac{SAPT_2 \cap SAPT_5}{SAPT_2 \cup SAPT_5} = \frac{3}{8}$ = 0.375, $SAPT_2 \cap SAPT_6$={root, A, AC}
$SAPT_2 \cup SAPT_6$= {root, A, B, C, AC, CB, ACB, AD}
Normalized similarity measure between trees $SAPT_2$ and $SAPT_6$ =$\frac{SAPT_2 \cap SAPT_6}{SAPT_2 \cup SAPT_6} = \frac{3}{8}$ =0.37
$SAPT_3 \cap SAPT_4$= {root, B, BC, BD}
$SAPT_3 \cup SAPT_4$= {root, B, C, D, BC, BD, CD}
Normalized similarity measure between trees $SAPT_3$ and $SAPT_4$ =$\frac{SAPT_3 \cap SAPT_4}{SAPT_3 \cup SAPT_4} = \frac{4}{7}$ =0.57, $SAPT_3 \cap SAPT_5$= {root}
$SAPT_3 \cup SAPT_5$= { root, B, BC, BD, A, AB, AC }
Normalized similarity measure between trees $SAPT_3$ and $SAPT_5$ =$\frac{SAPT_3 \cap SAPT_5}{SAPT_3 \cup SAPT_5} = \frac{1}{7}$ =0.15, $SAPT_3 \cap SAPT_6$= {root}
$SAPT_3 \cup SAPT_6$= {root, B, BC, BD, A, AC, AD}

Normalized similarity measure between trees $SAPT_3$ and $SAPT_6 = \frac{SAPT_3 \cap SAPT_6}{SAPT_3 \cup SAPT_6} = \frac{1}{7} = 0.15$, $SAPT_4 \cap SAPT_5 = \{root\}$
$SPT_4 \cup SPT_5 = \{root, B, C, D, BC, BD, CD, A, AB, AC\}$
Normalized similarity measure between trees $SAPT_4$ and $SAPT_5 = \frac{SAPT_4 \cap SAPT_5}{SAPT_4 \cup SAPT_5} = \frac{1}{10} = 0.1$, $SAPT_4 \cap SAPT_6 = \{root\}$
$SAPT_4 \cup SAPT_6 = \{root, B,C,D, BC,BD,CD,A, AC, AD\}$
Normalized similarity measure between trees $SAPT_4$ and $SAPT_6 = \frac{SAPT_4 \cap SAPT_6}{SAPT_4 \cup SAPT_6} = \frac{1}{10} = 0.1$
$SAPT_5 \cap SAPT_6 = \{root, A, AC\}$
$SAPT_5 \cup SAPT_6 = \{root, A, AB, AC, AD\}$
Normalized similarity measure between trees $SAPT_5$ and $SAPT_6 = \frac{SAPT_5 \cap SPT_6}{SAPT_5 \cup SPT_6} = \frac{3}{5} = 0.6$
Based on the highest similarity measure values sequential probability tree $SAPT_5$ and $SAPT_6$ are clustered, and then $SAPT_3$and $T_4$are clustered and then $SAPT_1$and $SAPT_2$are finally clustered. Among all these formal clusters the same process can be repeated to construct larger clusters.

6. Conclusion and Future Work

A trajectory is a useful tool to study a pattern of any situation. Mining trajectory data gives us useful information on movement patterns. This pattern information is useful in analysis of the data and application of the knowledge mined towards decision making. The route information of a public transportation system is considered for analysis here can be used in future to implement a good decision support system. The clustering technique proposed here can be applied on the data relating to various situations of the business. The same can be used on any transportation data like airlines, logistics; etc. The information obtained can be applied to improve customer relations and the business span as well. The market basket data relating to a single customer or a single customer group can be analyzed and the customer centric clusters can be formed to target specific customer groups. The present work is concerned with one similarity clustering measure in detail. In future there is a scope to deal with so many such measures. A great prospect is there to apply the same to reinforce the CRM practices.

References

[1] Ali Shahbazi, Student Member, IEEE and James Miller, Member, IEEE "Extended Subtree: A New Similarity Function for Tree Structured Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4, APRIL 2014

[2] Ayman A1-serafi, Teradata Corporation and Ahmed Elragal, Department of Business Informatics and Operations German university in cairo, "GEO Processing 2015", The Seventh International Conference on Advanced Geographic Information Systems, Applications, and Services ISBN: 978-1-61208-383-4 February 22 - 27, 2015

[3] Babita Chopra ,Vivek Bhambri and Balram Krishan, " Implementation of Data Mining Techniques for Strategic CRM Issues", Vivek Bhambri et al, Int. J. Comp. Tech. Appl., Vol 2 879-883, IJCTA | JULY-AUGUST 2011 , ISSN:2229-6093.

[4] Chris Rygielski , Jyun-Cheng Wang , and David C. Yen a, Chung-Cheng University, Taiwan, "Data mining techniques for Customer Relationship Management ", ROC Technology in Society 24 (2002) 483–502.

[5] Gaurav Gupta and Himanshu Aggarwal "Improving Customer Relationship Management using Data Mining " International Journal of Machine Learning and Computing, Vol. 2, No. 6, December 2012.

[6] Jean Damascène Mazimpaka, and Sabine Timpf " Trajectory data mining - A review of methods and applications", JOURNAL OF SPATIAL INFORMATION SCIENCE, 2016.

[7] Longbing Cao, Philip S.Yu, Chengqi Zhang, Huaifeng Zhang "Data Mining for Business Applications", Springer text book.

[8] Luis Otavio Alvares Vania Bogorny Jose Antonio Fernandes de Macedo Bart Moelans Stefano Spaccapietra "Dynamic Modeling of Trajectory Patterns using Data Mining and Reverse Engineering " 26th International conference on Conceptual Modeling- ER 2007, CRPIT , volume 83, Theoretical Computer Science Group, Hasselt University, Belgium.

[9] Petter Kihlstrom , "Literature Study and Assessment of Trajectory Data Mining Tools" Degree Project in Built Environment, First Cycle STOCKHOLM 2015.

[10] Sakshi Sivarama Krishna, Cheruku Sudarsana Reddy "E-Customer Classification using Data Mining for CRM" International Journal of Engineering Research & Technology (IJERT) NCACI-2015 Conference Proceedings.

[11] Sunil Yadav , Aaditya Desai and Vandana Yadav "Knowledge Management in CRM using Data mining Technique" ,International Journal of Scientific & Engineering Research, Volume 4, Issue 7, July-2013 465 ISSN 2229-5518 IJSER © 2013 http://www.ijser.org .

[12] Susanta Satpathy, Lokesh Sharma, Ajaya K. Akasapu, Netreshwari Sharma "Towards Mining Approaches for Trajectory Data" International Journal of Advances in Science and Technology Vol. 2, No.3, 2011

[13] Tipawan Silwattananusarn, Dr. KulthidaTuamsuk, Assoc.Professor, "Data Mining and its Applications for Knowledge Management" : A Literature Review from 2007 to 2012 .

[14] Uma Maheswari . R, Saravana Mahesan. S, Dr. Tamilarasan , A. K. Subramani , "Role of Data Mining in CRM" International Journal of Engineering Research (ISSN:2319-6890)(online),2347-5013(print) Volume No.3, Issue No.2, pp : 75-78 01 Feb. 2014.

[15] Vania Bogorny, Carlos Alberto Heuser, and Luis Otavio Alvares "A Conceptual Data Model for Trajectory Data Mining",Geographic Information Science, Lecture Notes in Computer Science, Volume 6292. ISBN 978-3-642-15299-3. Springer-Verlag Berlin Heidelberg, 2010, p. 1, DOI:10.1007/978-3-642-153006-6_1, Source DBLP

[16] Wen-Yuan Zhu, Wen- Chih Peng, Member, IEEE, Chih-Chieh Hung, Po-Ruey Lei, and Ling-Jyh Chen, Senior Member, IEEE, "Exploring Sequential Probability Tree for

Movement- Based Community Discovery", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 11, NOVEMBER 2014

[17] Xin Cao, G. Cong, and C.S. Jensen, "Mining Significant Semantic Locations from GPS Data," Proc. VLDB Endowment,vol.3, no.1, pp. 1009-1020, Sept. 2010.

[18] YU ZHENG. "Trajectory Data Mining: An Overview" Microsoft Research, ACM Transactions on Intelligent Systems and Technology, Vol 6, No.3, Article 29, May 2015.

[19] Zheng Y. 2015. "Trajectory Data Mining: An Overview", ACM Transactions on Intelligent Systems and Technology (TIST) - Survey Paper, Regular Papers and Special Section on Participatory Sensing and Crowd Intelligence, Volume 6 Issue 3, May 2015 Article No. 29.

[20] ZHENNI, FENG, AND YANMIN ZHU(Member IEEE), Department of Computer Sciencee & Engineering, Shanghai Jiao Tong University, "A Survey on Trajectory Data Mining: Techniques and Applications", Volume 4, April, 2016 , Digital Object Identifier 10.1109/ACCESS.2016.2553681.

[21] Zhixian Yan, supervised by Prof. Stefano Spaccapietra, EPFL Swiss,Federal Institute of Technology, &Lausanne, Switzerland, "Towards Semantic Trajectory Data Analysis: A Conceptual and Computational Approach", VLDB 09, August-24-28, 2009, Lyon, France, ACM

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. She is a research scholar in the department of Computer Science, Sri Venkateswara University, Tirupati, AP, India. Her research focus is on Applications of Data Mining in Customer Relationship Management. .

P. GOVINDARAJULU, 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