ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# Trajectory Clustering Approaches for Location Aware Services

Sanjiv Kumar Shukla
Department of Computer Science and Engineering
Rungta College of Engineering and Technology
Bhilai, India
sanjeev@rungta.ac.in

Sourabh Rungta
Department of Computer Science and Engineering
Rungta College of Engineering and Technology
Bhilai, India
sourabh@rungta.ac.in

Lokesh Kumar Sharma*
Department of Computer Science and Engineering
Rungta College of Engineering and Technology
Bhilai, India
lksharmain@gmail.com

***Abstract:*** Modern Geographic Information Systems (GIS) can handle dynamically moving objects. It is becoming possible to record data about the movement of people and objects at a large scale. Location Aware Service (LAS) is service as the application of which the service and information provided is determined by the user location. In this study, we consider the trajectory data to identify the important location for services. Clustering algorithms for these trajectory objects provide new and helpful information, e.g. location aware services, traffic jam detection and identifying the interest place. Trajectory data contains uncertain positional information. In this paper, DBSCAN, k-means and Fuzzy C Mean clustering approach for trajectory data is implemented and tested and in real dataset. The results are reported in this study.

***Keywords:*** Location Aware Services; Trajectory Cluster; DBSCAN; Fuzzy C-Means; k-Means.

## I. INTRODUCTION

Modern location-aware devices and applications deliver huge quantities of spatiotemporal data of moving objects, which must be either quickly processed for real-time applications, like traffic control management, or carefully mined for complex, knowledge discovering tasks. The analysis of mobile behavior leads to many instructive insights about the habits of a city's or a country's population. Always government and other organizations perform the study to evaluate mobility data with respect to travel distance, the means of transportation and the purpose of traveling. The mobility data contains both spatial and temporal features. Spatiotemporal data is typically stored in 3-D format (2-D space information + time). Thus traditional data mining algorithms cannot be applied in their original form. Therefore, knowledge discovery from mobility data has become an important research area [1] [2].

There is a clear and present need to bring together researchers from academia, government, and the private sector in the broad areas of knowledge discovery from trajectory data such as trajectory data pre-processing, representation and transformation, scalable and distributed classification, prediction, clustering algorithms, space-time sampling techniques, etc. In this thesis we will develop some data mining techniques such as classification and clustering for trajectory data [3].

Mobility data mining is, therefore, emerging as a novel area of research, aimed at the analysis of mobility data by means of appropriate patterns and models extracted by efficient algorithms; it also aims at creating a novel knowledge discovery, explicitly tailored to the analysis of mobility with reference to geography, at appropriate scales and granularity. In fact, movement always occurs in a given physical space, whose key semantic features are usually represented by geographical maps; as a consequence, the geographical background knowledge about a territory is always essential in understanding and analyzing mobility in such territory.

Mobility data mining, therefore, is situated in a Geographic Knowledge Discovery process – a term first introduced by Han and Miller in [4] – capable of sustaining the entire chain of production from raw mobility data up to usable knowledge capable of supporting decision making in real applications. Figure 1 illustrated the trajectories of four entities moving over 20 times steps. The following patterns are highlighted: a flock of three entities over five times steps, a periodic pattern where an entity shows the same spatio-temporal pattern with some periodicity, a meeting place where three entities meet for four times steps, and finally, a frequently visited location, which is a region where a single entity spends a lot of times [5]. As a prototypical example, assume that source data are positioning logs from mobile cellular phones, reporting user's locations with reference to the cells in the GSM network. These mobility data come as streams of raw log entries recording as users entering a cell – (userID, time, cellID, in) – users exiting a cell – (userID, time, cellID, out) – or, in the near future, user's position within a cell – (userID, time, cellID, X, Y) and, in the case of GPS/Galileo equipped devices, user's absolute position. Indeed, each time a mobile phone is used on a given network, the phone company records real-time data about it, including time and cell location. If a call is taking place, the recording data-rate may be higher. Note that if the caller is moving, the call transfers seamlessly from one cell to the next. In this context, a novel geographic knowledge discovery process may be envisaged, composed of three main steps: trajectories reconstruction, knowledge extraction and delivery of the information obtained.

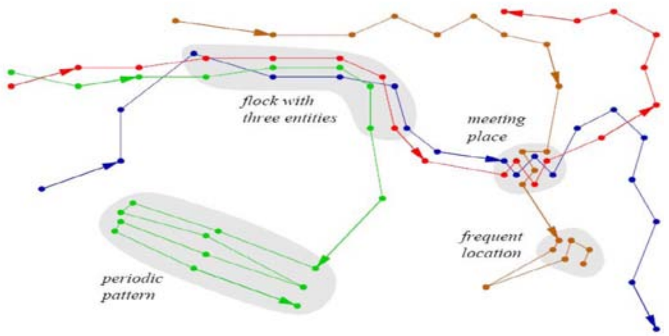


Figure 1: Some common movement patterns (Source [5]).

Trajectory clustering is an important research area in trajectory data mining [6] [7]. Clustering, the discovery of groups of ‘similar’ trajectories, together with a summary of each group. It is shown in Figure 2. Knowing which are the main routes (represented by clusters) followed by people or vehicles during the day can represent precious information for mobility analysis. For example, trajectory clusters may highlight the presence of important routes not adequately covered by the public transportation service.

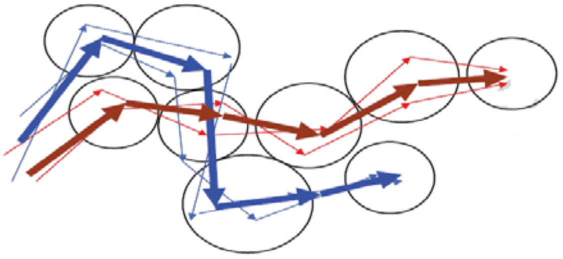


Figure 2: Illustration of Trajectory Clustering (Source:[1]).

In this study, the k-means, Fuzzy c-means and DBSCAN clustering algorithms are discuss to discover group of point or location and use to identify the interested point or locations. Thus, location aware services can be applied.

The remainder of this paper is organized as follows. In section 2, we will present the related work in the area of trajectory clustering. In section 3, we will k-means, Fuzzy C means and DBSCAN clustering techniques. Data pre-processing and result analysis is reported in section 4. Finally, the work is concluded in the section 5.

## II. RELATED WORKS

In this section, we will present the related work in the area of clustering moving objects. Given a set of objects with a distance function on them, an interesting data mining question is, whether these objects naturally form groups (called clusters) and what these groups look like. Data mining algorithms that try to answer this questions are called clustering algorithms.

### A. *Clustering Algorithms:*

Clustering algorithms can be classified in two ways. One way is according to the result they produce. For example, consider the hierarchical and partitioning clustering algorithms [8]. Partitioning algorithms construct a flat (single level) partition of a database D of n objects into a set of k clusters such that the objects in a cluster are more similar to each other than the objects in different clusters. Hierarchical algorithms decompose the database into several levels of nested partitioning (clustering), represented for example by a dendrogram, i.e. a tree that iteratively splits D into smaller subsets until each subset consists of only one object. In such a hierarchy, each node of the tree represents a cluster of D.

Another way classification is according to an algorithmic point of view. Here, we can distinguish between optimization based or distance based (k-means [8]) algorithms and density based (DBSCAN [9]) algorithms. Distance based methods use the distances between the objects directly in order to optimize a global criterion. In contrast, density based algorithms apply a local cluster criterion. Clusters are regarded as regions in the data space in which the objects are dense, and which are separated by regions of low object density (noise).

### B. *Fuzzy Clustering:*

Fuzzy clustering [10] arises as a commonly used conceptual and algorithmic framework for data analysis and unsupervised pattern recognition. In fact, fuzzy algorithms are an extension of the classical clustering algorithms to the fuzzy domain. However, there are very few efforts, in the field of fuzzy clustering that efficiently handle clusters of non-standard shapes.

A widely used fuzzy clustering algorithm is Fuzzy C-Means (FCM). It is an extension of the K-Means algorithm for fuzzy applications [10]. FCM attempts to find i) the representative point of each cluster, which is considered the “center” of the cluster, and ii) the degree of membership for each object to the defined clusters. It is obvious that FCM presents the similar disadvantages with K-Means. Considering that the fuzzy clusters are represented by their centers, the degree of data membership to a cluster decreases as the points move away from the center of a cluster. Thus, it mostly favors spherical clusters.

### C. *Clustering on Moving Objects:*

Yiu and Mamoulis [11] tackled the complex problem of clustering moving objects based on a spatial network. Here, the distance between the objects is defined by their shortest path distance over the network. Based on this distance measure they proposed variants of well-known clustering algorithms. Clustering moving objects is not only interesting in its own, but can also beneficially be used for spatio-temporal selectivity estimation [12].

## III. FRAME WORK OF CLUSTERING APPROACH FOR LAS

### A. *K-Means:*

The K-Means algorithm is the well-known clustering technique used in scientific and industrial applications. It is based on squared error criterion. K-mean algorithm initializes K-partition randomly and they change clusters based on their similarity between the objects and the cluster centroid C until a convergence criterion is met. K-Mean algorithm can be summarized as algorithm 1. The K-means algorithm is simple, relatively scalable and efficient and it can be easily implemented in solving many practical problems. The time complexity of K-mean algorithm is O (NKd) where d is number of iterations. Apart from these benefits, K-means algorithm has various loose falls [4] [8]. Therefore, several enhancements of the K-means algorithm have been reported. K-means algorithm deals only numerical data set. Due to lack of universal method for identification the initial number of partitions, the convergence centroids vary with different initial points. It is sensitive to noise and outlier data objects since a small number of data can influence the mean value.

```
Algorithm: 1 K-Mean Clustering Technique.
Input: The number of cluster K and a database containing N Objects.
Output: A set of K-clusters
Step 1. Initialize a K-partition randomly
Step 2. Calculate the initial centroids C for each K-cluster.
Step 3. For each object 'x'.
    a. Compute the dissimilarity between y and the centroids of all clusters
    b. Insert x into the cluster L whose centroid closest to x.
Step 4. Re-compute the cluster centroids so until the cluster dissimilarity between centroid and objects is minimized.
Step 5. Repeat 3 and 4 until no or few objects change clusters after a full cycle test of all the objects.
```

## *B. FuzzyC-Means:*

Let us also assume that the total no. of data sets are $N$ which is to be clustered in C groups, where $2 \leq C \leq N$ and $g(g>1)$ is a weighting factor indicating level of cluster fuzziness. Let us represent the membership matrix by [µ] and it has the dimension of N x C. Thus the membership value of ith data point with jth cluster is represented by $\mu_{ij}$. It is to be noted that $\mu_{ij}$ lies in the range of (0.0, 0.1). The Euclidean distance $d_{ij}$ is considered for dissimilarity measure which is the Euclidean distance between ith data point with jth cluster, which is calculated like the following Equation 1.

$$d_{ij} = \| c_i - x_j \| \qquad (1)$$

```
Algorithm 2: Fuzzy C-Means Clustering
Step 1. Assume the number of clusters to be made, that is, C where $2 \leq C \leq N$.
Step 2. Choose appropriate level of cluster fuzziness (g>1)
Step 3. Initialize the N x C size membership matrix [µ] at random such that $\mu_{ij}$ which lies in the range of (0.0, 0.1) and satisfy the condition below:
```

$$\sum_{ji=1}^{c} \mu_{ij} = 1.0 \text{ , for each } i. \qquad (2)$$

```
Step 4: Calculate Kth dimension of jth cluster center $CC_{jk}$ using the expression given below.
```

$$CC_{jk} = \frac{\sum_{i=1}^{N} \mu_{ij}^{g} x_{ik}}{\sum_{i=1}^{N} \mu_{ij}^{g}} \qquad (3)$$

```
Step 5. Calculate the Euclidean distance between ith data point and jth cluster center by Equation (2).
Step 6. Update fuzzy membership matrix [µ] according to $d_{ij}$. If $d_{ij} > 0$, then
```

$$\mu_{ij} = \frac{1}{\sum_{m=1}^{C} \left(\frac{d_{ij}}{d_{im}}\right)^{\frac{2}{g-1}}} \qquad (4)$$

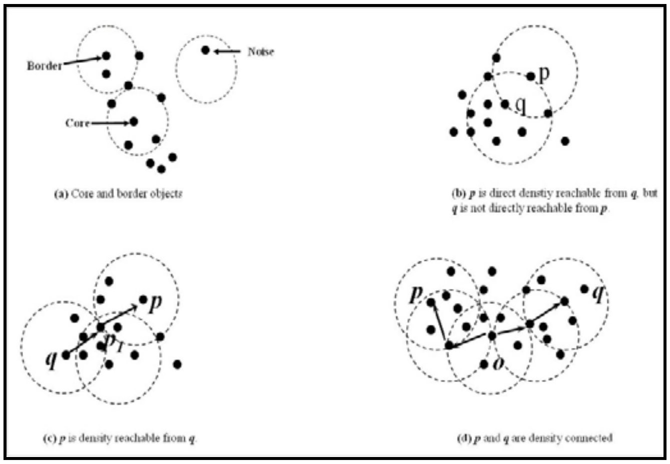
```
    If, $d_{ij} = 0$ then the data point coincides with jth cluster center ($CC_j$) and it will have the full membership values, that is , $\mu_{ij} = 1.0$.
Step 7. Repeat from step 4 to step 6 until the changes in[µ]come out to be less than pre-specified values. Using this algorithm, fuzzy clusters of the data set will be generated [10].
```

## *C. DBSCAN:*

Ester et al. [9] introduced density based algorithms DBSCAN. The key feature of DBSCAN (Density-Based Spatial Cluster of Applications with Noise) is that for each object of a cluster the neighbourhood of a given radius ε has to contain at least a specified minimum number MinC of objects, i.e., the cardinality of the neighbourhood has to exceed a given threshold. Radius ε and minimum number MinC of objects are specified by user. Let D is a data set of objects, the distance function between the objects of D is denoted by DIST and given parameters are ε and MinC then DBSCAN can be specified by the following definitions. We have adopted these definitions from Ester et al. [9].

Definition 1 (Neighbourhood of an object). The ε-neighbourhood of an object p, denoted by Nε (P) is defined by $N\varepsilon(P) = \{q \in D \mid DIST(p, q) \leq \varepsilon\}$.

Definition 2 (Direct Density Reachability). An object p is direct density reachablitiy from object q w.r.t. ε and MinC if $|N\varepsilon(P)| \geq MinC \wedge p \in N\varepsilon(q)$. q is called core object when the condition $|N\varepsilon(P)| \geq MinC$ holds (Figure 3 (a,b)).



(a) Core and border objects

(b) *p* is direct denstiy reachable from *q*, but *q* is not directly reachable from *p*.

(c) *p* is density reachable from *q*.

(d) *p* and *q* are density connected

Figure 3: Density based clustering concepts (MinC = 5).

Definition 2.3 (Density Reachability). An object p is density-reachable from an object q w.r.t ε and MinC if there is a sequence of objects $p_1 \ldots p_n$; $p_1 = q$, $p_n = p$ such that $p_{i+1}$ is directly density reachable for from pi (Figure 3 (c)).

DBSCAN chooses an arbitrary object p. It begins by performing a region query, which finds the neighbourhood of point q. If the neighbourhood contains less than MinC objects, then object p is classified as noise. Otherwise, a cluster is created and all objects in p's neighbourhood are placed in this cluster. Then the neighbourhood of each of p's neighbours is examined to see if it can be added to the cluster. If so, the process is repeated for every point in this neighbourhood, and so on. If a cluster cannot be expanded further, DBSCAN chooses another arbitrary unclassified object and repeats the same process. This procedure is iterated until all objects in the dataset have been placed in clusters or classified as noise.

## IV. EPERIMENT INVESTIGATION

The raw data of trajectory contains vehicle id, time stamp and location (longitude and latitude). Therefore raw data is required to preprocess before applying the clustering. In this section data preprocessing is reported and further the result analysis is discussed.

### *A. Data Preprocessing:*

For the task of clustering on trajectory data, we used Milan datasets. These data contain the records of moving vehicles in Milan City, Italy, which is provided by Milan Metropolitan Authority for research purpose. Data consists of positions of the vehicles, which has been GPS-tracked between April 1, 2007 and April 7, 2007, and are stored in a relational database. The data have been recorded only while the vehicles moved. Each record includes the vehicle-id, date and time, the latitude, longitude, and altitude of the position. To facilitate analysis of movement data, initial preprocessing in the database is performed, which enriches the data with additional fields: the time of the next position in the sequence, the time interval and the distance in space to the next position, speed, direction, acceleration (change of the speed), and turn (change of the direction) [13][14][15].

The most visited location is an important in term of location aware services. To identify the interest location, location aware service provider can provide the facility and plan the other marketing strategies. This regards the trajectory are split into several trajectories with respect to time. In this study we split trajectory on where vehicle is stopped between 30-60 minutes. We assume here the vehicle user stop for short period. These locations are not home and work place. We do not consider very short time stop such as 3-15 minutes. We assume this stop may be due to traffic signals or other. Further the ending points are identified and clustering algorithms are applied on those data points to perform the group of points or clusters.

### *B. Result Analysis:*

Above pre-processed data are mined using k-means, Fuzzy C Means and DBSCAN algorithm. The total 996 points are taken for experiments. The distributions of data are shown in figure 4, 5, and 6 respectively. The mobility data contain uncertainty therefore Fuzzy C means clustering is used here. The location can be on Density hence the DBSCAN is also used. The DBSCAN produces cluster and noise. The Noise is represented as diamond shape in figure. The DBSCAN produces the 14% noise in this study. The accuracy is measured by standard formula. These algorithms are suitable for performing clustering on location based data.

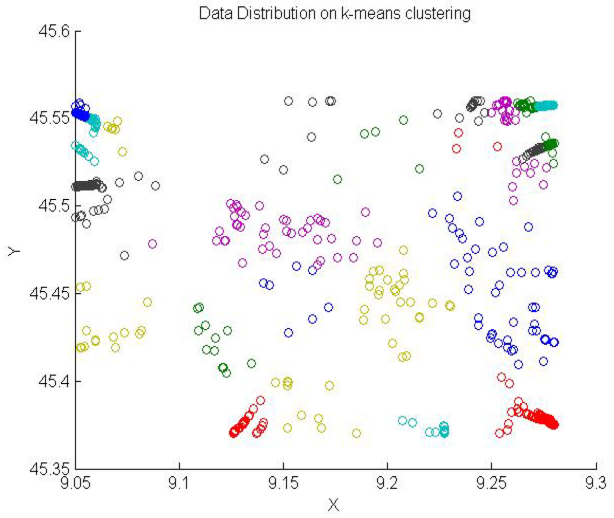


Figure 4: Cluster data by using k-means algorithm.

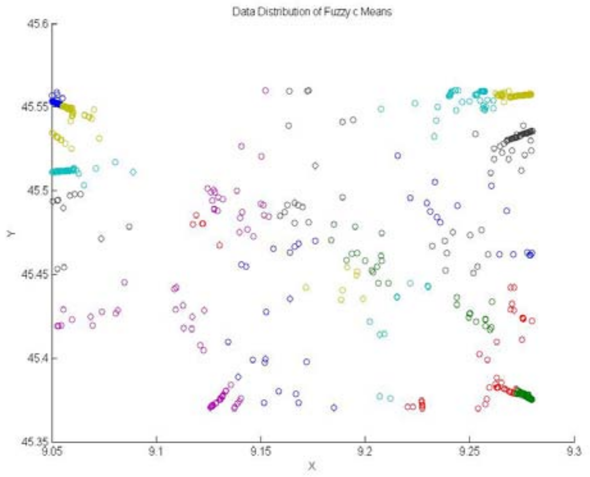


Figure 5: Cluster data by using Fuzzy c-means algorithm.

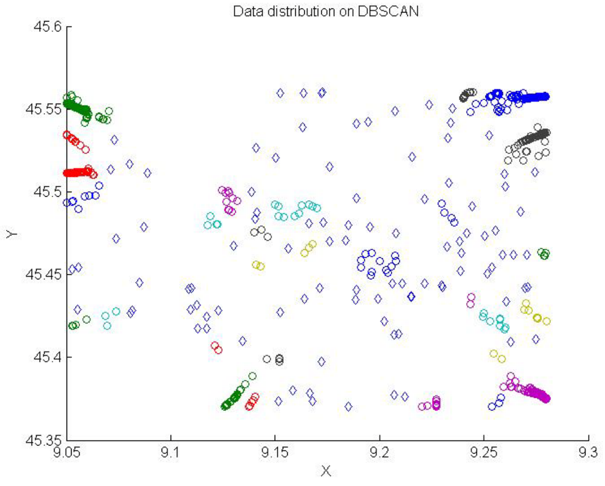


Figure 6: Cluster data by using DBSCAN algorithm.

## V. CONCLUSION

There is clearly an increasing demand for LAS applications and while the developed retype is basic in its current stage, it is able to identify the location of an information device user, search for offers that are within a defined range and present the offers to the users, the findings of this research have provided preliminary empirical evidence about how users are willing to strike a balance between value and risk. In this study, the trajectory clustering framework is provided in this regards. The trajectory clustering can be used as an important tool to identify the important location so that the services can be applied.

## VI. REFERENCES

[1] F. Giannotti and D. Pedreschi, "Mobility, Data Mining and Privacy: Geographic Knowledge Discovery", Springer Verlag, 2008.

[2] F. Ginnotti, M. Nanni, D. Pedreschi and F. Pinelli, "Trajectory Pattern Mining", In Proceedings of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 330 – 339, 2007.

[3] N. Pelekis, I Kopanakis, G. Marketos, I. Ntoutsi, G. Andrienko, and Y. Theodoridis , "Similarity Search in Trajectory Databases", In Proc. of 14th International Symposium on Temporal Representation and Reasoning, IEEE Computer Society, pp. 129-140, 2007

[4] J. Han and H. J. Miller, "Geographic Data Mining and Knowledge Discovery", CRC Press 2001.

[5] J. Gudmundsson, P. Laube and T. Wolle T. "Movement Patterns in Spatio-Temporal Data", In: Shekhar, S. & Xiong, H. (eds.). Encyclopedia of GIS, Springer-Verlag, 2008.

[6] J. Lee, J. Han., and K. Whang ,"Trajectory clustering: a partition-and-group framework", In Proceedings ACM SIGMOD Int. Conf. on Management of Data: 593 – 604, 2007.

[7] J. Lee J Han J, X. Li and H. Gonzalez, "TraClass: Trajectory classification Using Hierarchical Region Based and Trajectory based Clustering", In Proceedings ACM, VLDB, New Zealand: 1081-1094, 2008.

[8] A. K. Jain, M. N. Murty, P. J .Flynn, " Data Clustering: A Review." ACM Computing Surveys, Vol. 31, No. 3, pp. 265-323, Sep. 1999.

[9] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." KDD', pp. 226-231, 1996.

[10] F.Höppner ., F.Klawonn .,R. Kruse .,T. Runkler, "Fuzzy Cluster Analysis." Wiley, 1999.

[11] M.L.Yiu ., N. N.Mamoulis, " Clustering Objects on a Spatial Network." Proc. 23th ACM SIGMOD, pp. 443-454, 2004.

[12] Q. Zhang, X. Lin, "Clustering Moving Objects for Spatio-Temporal Selectivity Estimation." Proc 15th Australasian Database Conference (ADC), pp. 123-130, 2004.

[13] G. Andrienko, D. Malerba, M, May and M. Teisseire, "Mining spatio-temporal data", Journal of Intelligent Information Systems 27(3), pp. 187–190, 2006.

[14] G. Andrienko, N. Andrienko, and S. Wrobel, "Visual Analytics Tools for Analysis of Movement Data", ACM SIGKDD: 38-46, ISSN:1931-0145, 2007.

[15] L. K. Sharma,, O. P. Vyas, S. Scheider and A. Akasapu, "Nearest Neibhour Classification for Trajectory Data", ITC 2010, Springer LNCS CCIS 101, pp.180–185..