International Journal of Emerging Trends in Science and Technology
IC Value: 76.89 (Index Copernicus) Impact Factor: 4.219 DOI: https://dx.doi.org/10.18535/ijetst/v4i9.02

Clustering of Trajectory Data using Maximal Set Matched Method

Authors
Dr. S. Aquter Babu
Assistant Professor, Department of Computer Science
Dravidian University
Kuppam, Chittoor District, Andhra Pradesh, India

Abstract
Trajectory data is ubiquitous in many real time areas. Sequential collections of movements of objects is called trajectory. Trajectory data clustering is one of the most important and useful area of many modern trajectory data applications. Trajectory data represents the actual mobility of a diversity of dynamically moving objects, such as people, flights, birds, vehicles and animals. Many techniques have been proposed in the literature of trajectory data mining for processing, managing and mining trajectory data in the past. Most important tasks of trajectory data mining are - trajectory data preprocessing, trajectory data management, and also different varieties of trajectory data mining tasks such as trajectory pattern mining, outlier detection, and trajectory classification. Time complexity of many clustering data algorithms is $O(n^2)$. A new trajectory similarity measure is proposed for clustering trajectory data sets. The time complexity of proposed trajectory data clustering algorithm is O(n) which is very much better than time complexities of many real time clustering algorithms.

Keyword- Trajectory data mining, Trajectory data clustering, trajectory data similarity measure, intersection and union operations, maximal set matched.

1. Introduction

Nowadays trajectory data is very useful in optimal decision making in many real time applications. The sequential sequences of collected positions of objects such as users, animals, vehicles and flights are called trajectories. The sequences of these collected positions can be viewed as users' trajectories [1]. Trajectories represent actual movements of objects in the real life. Some of the most important trajectory operations are – trajectory classification, clustering, association, pattern mining, and outlier detection and so on. Trajectory clustering is mainly based on trajectory similarity finding measures. Trajectory clustering is mainly dependent on trajectory data features as well as multidimensional consideration of different features.

For trajectory data clustering one must consider similarities of objects in terms of their trajectories. Already existing trajectory similarity finding measures are – largest common subsequence (LCSS), Euclidian distance(ED), DTW, ERP, and EDR and so on. Trajectory details of objects are represented as trajectory data sets. These trajectory data sets are called trajectory profiles. Trajectory data sets are clustered using trajectory similarity measures.

2. Trajectory Pattern Mining

The goal of trajectory data pattern mining is to find frequent sequential patterns. Efficient data structures

are required to organize, store and efficient processing of these trajectory data patterns. Some of the important and efficient trajectory data structures are - trajectory indexing structure, trajectory pattern tree, T-pattern tree. The main requirement of trajectory pattern mining is to find movement behaviors of objects, such as users, vehicles, animals and so on. Total number of trajectory patterns increases as the number of trajectories increases. Trajectory clustering is a method of forming trajectories into groups called trajectory clusters.

Trajectories within the group have similar features and trajectories in the other groups have dissimilar features. Trajectory pattern mining aims to discover frequent sequential patterns that are sequential relationships among regions [1]. The state-of-the-art trajectory data clustering approach is TraClus [2]. Cosine similarity and LCSS (longest common sub sequence), and Euclidean methods are some of the popular trajectory similarity measures, the similarity function, HGSM is designed for comparing the similarity between two users when trajectory data is implemented for performance comparison [3].

Trajectories belonging to one user are efficiently represented in the form of tree data structure and then these trees may be clustered using best clustering algorithm.

3. Problem Definition

A trajectory is a set of sequential locations of moving objects. In general, a trajectory is represented as (location-1, location-2, location-3, …, location-n). Movement details of objects are considered and used with respect to multidimensional view of moving objects. Prior works [4] have elaborated on discovering user communities from user location history. Raw trajectory data details must be translated into modified, convenient, and useful format for their efficient and effective processing in their applications. Note that for similarity purpose trajectory is usually represented as

User-1 trajectories are-

$l_1^i, l_2^i, l_3^i, l_4^i, \ldots . , l_n^i$

$l_2^i, l_5^i, l_9^i, l_{12}^i, \ldots . , l_{n+5}^i$

$l_1^i, l_3^i, l_4^i, l_9^i, \ldots . , l_{n+7}^i$

……..

……..

User-2 trajectories are-

$l_1^j, l_2^j, l_3^j, l_4^j, \ldots . , l_n^j$

$l_2^j, l_5^j, l_9^j, l_{12}^j, \ldots . , l_{n+5}^j$

$l_2^j, l_3^j, l_5^j, l_9^j, \ldots . , l_{n+2}^j$

……..

……..

Some of the real time applications of trajectory data similarity measures usage examples are - Trajectory ranking, Community-based traffic sharing services and Friend recommendation [5]. Trajectories are also useful in city management systems.

A trajectory data set is a group of trajectories belonging to one user. N-number of trajectory data sets represents trajectories of N-users. Among the many possible and useful data mining techniques trajectory data clustering is most important technique and widely applicable technique in trajectory data mining. Trajectory data sets are generally clustered based on their trajectory similarity measures. Main steps in trajectory data clustering are –

1. Collect location details of trajectories.
2. Modify these locations so that modified locations are convenient and easy to use.
3. Store all these ‘n’ number of trajectory data sets corresponding to ‘n’ number of users in convenient data structures.
4. Use efficient trajectory data clustering algorithm to cluster all these trajectory data sets by using suitable trajectory similarity finding measure.

4. Proposed Trajectory Data Clustering Algorithm

Newly proposed trajectory data clustering algorithm is very useful in many real time applications. It

works based on trajectory similarity measure called intersection of trajectories divided by union of trajectories between two different trajectories. This measure is evaluated iteratively among the trajectories and based on this measure trajectories are clustered.

Algorithm EfficientTrajectoryClustering

1. Input trajectory profiles of 'n' number of users
2. Input threshold for trajectory data clustering
3. clusterCount = n * 0.25
4. while (n > clusterCount) Do
5. {
6. leadeCluster = 1
   /* first cluster of the ungrouped cluster */
7. find the trajectory similarity measures between
   leaderCluster and all the remaining clusters which
   are not grouped
8. store all the similarity measure in an array
9. Combine all the clusters whose similarity
   measures > threshold value into one group
10. k=number of clusters of the present group
11. n = n – k
12. }
13. Display all the groups of clusters

Description of effective trajectory clustering algorithm

The new trajectory similarity finding measure is used for clustering trajectories. This measure is denoted as by the symbol,

$$\nabla = \frac{\text{intersetion of two given trajectories}}{\text{union of two given trajectories}}$$

$$= T_1 \nabla T_2 = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

Where T_1 and T_2 are two different trajectory sets belonging to two different users, vehicles, animals and objectes and so on.

Initially 'n' number of trajectory sets corresponding to 'n' users are represented and stored in a convenient data structure. A suitable threshold for trajectory clustering is selected cleverly. If the data set contains small number of trajectory sets then select 25%; otherwise if data size is very large select 10% of total number of initial clusters as cluster count. While loop repeatedly executed until 'n' becomes less than cluster Count.

5. Example Data Set of Trajectory Data Clustering

Within the each iteration of the while loop the first cluster which is not grouped is selected as the leader Cluster. Trajectory similarity measures between the leader cluster and all the remaining ungrouped clusters are computed separately for each pair. For example, assume that initially 10 clusters are there. First cluster is taken as leader cluster (T_1). Now trajectory similarity measures between pairs (T_1, T_2), (T_1, T_3), …. , (T_1, T_{10}) are computed. If T_3, T_6 and T_9 satisfy the greater than threshold property then T_1, T_3 T6 and T_9 are grouped. Now remaining clusters are T_2, T_4, T_5, T_7, T_8, T_{10}.

In the second iteration of the while loop T_2 is selected as the leader cluster and the same process is repeated. That is trajectory similarity measures between pairs (T_2, T_4), (T_2, T_5), …. , (T_2, T_8) and (T_2, T_{10}) are computed. Again assume that (T_2, T_5) and (T_2, T_8) satisfy the threshold value and hence T_2, T_5, and T_8 are clustered. Now remaining initial clusters are T_4, T_7 and T_{10}. Same process is repeated for the remaining clusters. Here cluster count = 10 * 0.25 = ceiling of (2.5) = 3. While loop executes 3 times and are computed for each iteration of the while loop.

Consider the following user profiles of trajectories:

T_1 = {A, B, C, AB, AC, BC, ABC}

T_2 = {B, C D, BC, BD, CD}

T_3 = {C, D, E, CD, CE, CDE}

T_4 = {E, F, G, H, EF, EG, EH, FGH}

T_5 = {A, B, C, AC, BC}

T_6 = {B, C, D, E, BC, BD, BE, CDE}

T_7 = {C, D, E, F, CD, CE}

T_8 = {E, F, G, H, EH, FGH}

T_9 = {A, B, C, AB, BC, ABC}

T_{10} = {B, C, D, E, BE, BC, CD, DE}

T_{11} = {C, D, E, F, CD, CE, EF}

T_{12} = {C, D, E, F, CE, CF, EF}

Above twelve trajectory sets are clustered using a new trajectory similarity measure called set-intersection/set-union. Initially assume that all the trajectory sets are individual clusters and T_1 is considered as leader cluster. Now trajectory similarity measures between T_1 and T_2, T_1 and T_3, ….T_1 and T_{12} are computed using proposed trajectory similarity measure. Trajectory similarity measure computations are shown below:

$$T_1 \nabla T_2 = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} = \frac{\{B, C, BC\}}{\{A, B, C, D, AB, AC, BC, BD, CD, ABC\}} = \frac{3}{10} = 0.3$$

$$T_1 \nabla T_3 = \frac{|T_1 \cap T_3|}{|T_1 \cup T_3|} = \frac{\{C\}}{\{A, B, C, AB, AC, BC, ABC, D, E, CD, CE, CDE\}} = \frac{1}{12} = 0.08$$

$$T_1 \nabla T_4 = \frac{|T_1 \cap T_4|}{|T_1 \cup T_4|} = 0$$

$$T_1 \nabla T_5 = \frac{|T_1 \cap T_5|}{|T_1 \cup T_5|} = \frac{\{B, C, BC\}}{\{A, B, C, AB, AC, BC, ABC\}} = \frac{5}{7} = 0.712$$

$$T_1 \nabla T_6 = \frac{|T_1 \cap T_6|}{|T_1 \cup T_6|} = \frac{\{A, B, C, AC, BC\}}{\{A, B, C, AB, AC, BC, ABC, D, E, BD, BF, CDF\}} = \frac{3}{13} = 0.23$$

$$T_1 \nabla T_7 = \frac{|T_1 \cap T_7|}{|T_1 \cup T_7|} = \frac{\{C\}}{\{A, B, C, AB, AC, BC, ABC, D, E, F, CD, CE\}} = \frac{1}{12} = 0.08$$

$$T_1 \nabla T_8 = \frac{|T_1 \cap T_8|}{|T_1 \cup T_8|} = 0$$

$$T_1 \nabla T_9 = \frac{|T_1 \cap T_9|}{|T_1 \cup T_9|} = \frac{\{A, B, C, AB, BC, ABC\}}{\{A, B, C, AB, AC, BC, ABC\}} = \frac{6}{7} = 0.8$$

$$T_1 \nabla T_{10} = \frac{|T_1 \cap T_{10}|}{|T_1 \cup T_{10}|} = \frac{\{B, C, BC\}}{\{A, B, C, AB, AC, BC, ABC, D, E, BE, CD, DE\}} = \frac{3}{12} = 0.25$$

$$T_1 \nabla T_{11} = \frac{|T_1 \cap T_{11}|}{|T_1 \cup T_{11}|} = \frac{1}{13} = 0.07$$

$$T_1 \nabla T_{12} = \frac{|T_1 \cap T_{12}|}{|T_1 \cup T_{12}|} = \frac{1}{13} = 0.07$$

Here maximal set matched is 4. Assume that threshold value set is 0.3 and as a result T_1, T_2, T_5 and T_9 are clustered. Hence, the resultant clusters are:

T_3 = {C, D, E, CD, CE, CDE}

T_4 = {E, F, G, H, EF, EG, EH, FGH}

T_5 = {A, B, C, AC, BC}

T_6 = {B, C, D, E, BC, BD, BE, CDE}

T_7 = {C, D, E, F, CD, CE}

T_8 = {E, F, G, H, EH, FGH}

T_{10} = {B, C, D, E, BE, BC, CD, DE}

T_{11} = {C, D, E, F, CD, CE, EF}

T_{12} = {C, D, E, F, CE, CF, EF}

Now trajectory similarity between T_3 and T_4, T_6, T_7, T_8, T_{10}, T_{11}, and T_{12} respectively are computed as follows:

$$T_3 \nabla T_4 = \frac{|T_3 \cap T_4|}{|T_3 \cup T_4|}$$
$$= \frac{\{E\}}{\{C, D, E, CD, CE, CDE, F, G, H, EF, EG, EH, FGH\}}$$
$$= \frac{1}{13} = 0.07$$

$$T_3 \nabla T_6 = \frac{|T_3 \cap T_6|}{|T_3 \cup T_6|}$$
$$= \frac{\{C, D, E, CDE\}}{\{C, D, E, CD, CE, CDE, B, BC, BD, BE\}} = \frac{4}{10}$$
$$= 0.4$$

$$T_3 \nabla T_7 = \frac{|T_3 \cap T_7|}{|T_3 \cup T_7|} = \frac{\{C, D, E, CD, CE\}}{\{C, D, E, CD, CE, CDE, F\}} = \frac{5}{7}$$
$$= 0.7$$

$$T_3 \nabla T_8 = \frac{|T_3 \cap T_8|}{|T_3 \cup T_8|}$$
$$= \frac{\{E\}}{\{C, D, E, CD, CE, CDE, F, G, H, EH, FGH\}} = \frac{1}{11}$$
$$= 0.1$$

$$T_3 \nabla T_{10} = \frac{|T_3 \cap T_{10}|}{|T_3 \cup T_{10}|}$$
$$= \frac{\{C, D, E, CE\}}{\{C, D, E, CD, CE, CDE, B, BE, BC, CD\}} = \frac{4}{10} = 0.4$$

$$T_3 \nabla T_{11} = \frac{|T_3 \cap T_{11}|}{|T_3 \cup T_{11}|} = \frac{\{C, D, E, CD, CE\}}{\{C, D, E, CD, CE, CDE, F, EF\}}$$
$$= \frac{5}{7} = 0.6$$

$$T_3 \nabla T_{12} = \frac{|T_3 \cap T_{12}|}{|T_3 \cup T_{12}|}$$
$$= \frac{\{C, D, E, CE\}}{\{C, D, E, CD, CE, CDE, F, CF, EF\}}$$
$$= \frac{4}{9} = 0.44$$

Here maximal set matched value = 6.

Now T_3, T_6, T_7, T_{10}, T_{11}, and T_{12} are clustered as threshold values are greater than 0.3. The remaining clusters are:

T_4 = {E, F, G, H, EF, EG, EH, FGH}

T_8 = {E, F, G, H, EH, FGH}

Now trajectory similarity measure between T_4 and T_8 are computed

$$T_4 \nabla T_8 = \frac{|T_3 \cap T_{12}|}{|T_3 \cup T_{12}|}$$
$$= \frac{\{E, F, G, H, EH, FGH\}}{\{E, F, G, H, E, F, EG, EH, FGH\}}$$
$$= \frac{6}{9} = 0.666$$

Here maximal set matched value = 2 and T_4 and T_8 are clustered.

The entire given trajectory sets are clustered which are shown below:

Cluster-1 = {T_1, T_2 , T_5 and T_9 }

Cluster-2 = {T_3, T_6, T_7, T_{10}, T_{11}, and T_{12} }

Cluster-3 = {T_4 and T_8}

5. TIME COMPLEXITY OF PROPOSED TRAJECTORY DATA CLUSTERING ALGORITHM

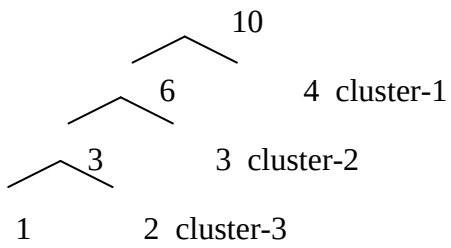

FIGURE-1

In this example time complexity of clustering algorithm is proved to be linear. In all the iterations number of comparisons for clustering 10 trajectory data sets are computed = (10 – 1) + (6 – 1) + (3 –1)

= 9 + 5 + 2 = 16

16 < 2 * 10

i.e n < 2 * n = O(n).

Assume that initially there are n = 10 separate individual simple clusters. For understanding purpose only a small data set is considered. In reality the sizes of trajectory data sets is very large. In FIGURE-1 ten trajectories are there. In the first iteration 4 trajectories are grouped into one cluster. This requires 9 comparisons. In the second pass in

the remaining 6 clusters again 3 are grouped into one cluster. In the next iteration 2 are grouped into another cluster and so on. Clusters are shown individually in tree data structure called cluster tree which shows number of clusters formed in sequential iterations of the clustering algorithm.

Also assume that now there are n = 20 separate individual simple clusters.

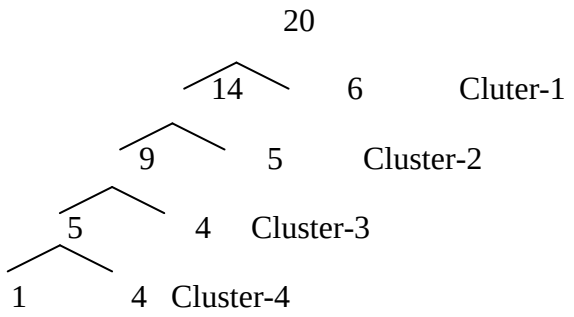

FIGURE-2

Total number of comparisons for clustering 20 trajectory data sets are computed = (20 – 1) + (14 – 1) + (9 –1) + (5 –1)

= 19 + 13 + 8 + 4

= 44

44 < 3 * 20

i.e. n < 3 * n = O(n)

Hence time complexity = O(n).

In the general case consider ‘n’n sets of trajectories. Also assume that now there are separate individual simple clusters.

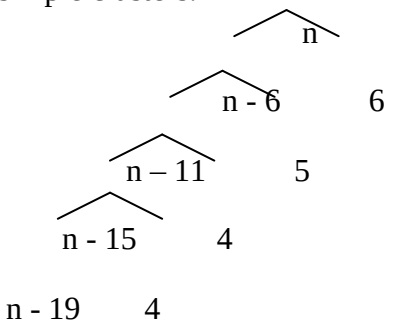

FIGURE-3 general value of n for forming clusters

Total number of comparisons in this case is

= (n – 1) + (n – 7) + (n – 12) + (n – 16) + (n – 20)

= 5*n – (1 + 7 + 12 + 16 + 20)

= 5n – 56 approximately = O(n).

When n = 20 the value of 5n – 56 = 5 * 20 – 56 = 44

Time complexity of newly proposed trajectory data clustering algorithm is O(n) which is far better than the normally existing general time complexity of O(n log(n) and $O(n^2)$ of many existing clustering algorithms. The difference will become very clear when data size is very large.

Conclusions

Efficient trajectory data clustering algorithm is proposed in this paper. This clustering algorithm is based on the new trajectory data similarity finding measure. Time complexity of proposed trajectory data clustering algorithm is O(n) where n is the number of trajectory data sets corresponding to ‘n’ number of objects (users, vehicles, animals and so on). In the future there is a scope for finding different types efficient trajectory similarity measures for trajectory data clustering and all these measures are have reasonably optimal time complexities of different types of trajectory similarity finding measures. Also, there is a scope for trajectory clustering based on multidimensional features instead of general consideration and usage of only one dimension details normally. Multidimensional consideration of trajectory data operations, definitely improves quality as well as accuracy of trajectory data clustering in many real time applications including special database applications.

References

1. 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 IEEETRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 11, NOVEMBER 2014

2. J.-G. Lee, J. Han, and K.-Y. Whang, “Trajectory Clustering: A Partition-and-Group Framework,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’07), June 2007.

3. Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining User Similarity Based on Location History," Proc. 16th
4. ACM SIGSPATIAL Int'l Conf. Advances in Geographic Information Systems (GIS '08), Nov. 2008.

5. X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Finding Similar Users using Category-Based Location History," Proc. ACM 18th
6. SIGSPATIAL Int'l Conf. Advances in Geographic Information Systems (GIS '10), Nov. 2010.

7. Y. Zheng and X. Zhou, Computing with Spatial Trajectories, first ed. Springer, 2011.