Detection of Flock Movement in Spatio-Temporal Database using Clustering Techniques - an Experience

Geethu Miriam Jacob
Department of Computer Science, CUSAT
geethumiriam@gmail.com

Sumam Mary Idicula
Cochin University of Science and Technology
sumam@cusat.ac.in

Abstract

In this paper, moving flock patterns are mined from spatio- temporal datasets by incorporating a clustering algorithm. A flock is defined as the set of data that move together for a certain continuous amount of time. Finding out moving flock patterns using clustering algorithms is a potential method to find out frequent patterns of movement in large trajectory datasets. In this approach, SPatial clusteRing algoRithm thrOugh sWarm intelligence (SPARROW) is the clustering algorithm used. The advantage of using SPARROW algorithm is that it can effectively discover clusters of widely varying sizes and shapes from large databases. Variations of the proposed method are addressed and also the experimental results show that the problem of scalability and duplicate pattern formation is addressed. This method also reduces the number of patterns produced.

Keywords: spatio-temporal data, flock patterns, clustering, frequent pattern mining

1. Introduction

Modern world has been conquered by new data acquisition techniques like Global Positioning Systems (GPS), Radio Frequency Identification (RFID) and wireless sensor networks. Many surveys have also been done using these techniques. All these have resulted in the tremendous increase in the amount of geographic data during the past few years. The current trends in the use of these mobile devices indicate that the amount of geographic data will increase in the near future.

Since the development of the data acquisition techniques, geo-referenced data are now increasing tremendously. Moving point object data is also becoming available in large numbers. One of the main objectives of spatial data mining is to analyze the data for interesting patterns. Initially spatial data of the movement of animals with the help of radio collars were analyzed to identify the migration patterns [1]. Such analysis helps in the identification of interaction between the entities. The group of entities moving together for a considerable amount of time is known as flocks [2]. The analysis of flocks can be extended to fields like socio- economic geography, transport analysis, and surveillance areas.

Early approaches of study of the moving trajectory datasets included the use of predicate and nearest neighbor queries. Questions like "how many cars drove from Main Square to Airport on Friday" were answered. Recently studies have been based on the group behavior of entities. The characteristics of the entities (animals, pedestrians, vehicles), how they interact with each other causes the flock pattern detection particularly relevant. The existing method considers group of trajectories within a circular radius for a certain amount of time as flocks [3]. The existing approach finds disks for each time instance and merges the results to get the patterns from one time instance to the other. The final patterns and the performance depend on the type and number of disks produced.

Clustering can be used for finding out flocks. Clustering can overcome the overlapping problem and can avoid the creation of duplicate patterns. Clustering can be defined as the division of data into distinct groups, with members within each group being similar to the members within that group but different from the members of other groups [4]. The clustering algorithm used in this problem is SPARROW because of the production of clusters in a decentralized fashion.

Frequent pattern discovery is also a research area in data mining. Association rule mining and frequent pattern mining are well researched areas to discover interesting relations between variables in large databases. Frequent patterns can be item sets or subsequences appearing in a datasets with frequency greater than user specified threshold. Frequent pattern

discovery can be used in finding moving flock patterns to get the interesting movement patterns.

2. Related works

Flocks are those trajectory data that stay together for a certain continuous amount of time. An existing method finds out flocks from spatio-temporal data by grouping geographical data falling within a circular radius for certain amount of time [3]. Since circular disk shapes are used for the grouping, there is a possibility that the same data is overlapped in more than one disk. This caused redundancy in the pattern formation. Spatial clustering algorithms are alternative to find out flocks of varying shapes and sizes instead of disks. This ensures that no data is included in more than one cluster. DBScan algorithm is a popular density based clustering algorithm. The algorithm uses parameters like minimum number of points (minpts) and maximum distance (ε) to contain in a cluster. But DBscan has problems with handling large databases and the worst case complexity of dBscan is $O(n^2)$ [5].

Recently an algorithm named SPARROW(Spatial Clustering algorithm through Swarm Intelligence) combining the concepts of DBscan and swarm intelligence techniques was introduced [6]. The algorithm combined the good aspects of DBscan algorithm and also could be extended to large databases.

In the field of finding out group behaviors, one of the first projects was finding out the migration patterns of moose in Sweden [1]. By knowing the movement pattern in different seasons, the hunters and foresters could move accordingly. Also, another reference on the movement of flocks of icebergs was given knowing which helped in creating awareness on global warming and climatic change. The movement of icebergs directly affected the growth of underwater diversity [7].

Predestination or predicting the destination based on the driver's behaviors, past movements, etc is a method which helps in location based services [8]. This was a Microsoft research result and the dataset [9] used for this research was used in this experiment.

In this paper, moving flock patterns are mined from spatiotemporal datasets using the SPARROW clustering algorithm. First, trajectory datasets are divided into time frames, i.e. the data belonging to each time frame is found out and separated. Secondly, the flocks are identified in each time frame using clustering. Third, a transactional version of the dataset is developed with the help of the clusters formed. Fourthly, associate rule mining algorithm is applied to the transactional version of the dataset and finally after some post processing, visualized in Google earth or Open Jump.

3. Problem statement

Flock patterns are considered as movement trends which help in the study and analysis of group behaviors and interactions among the entities. The problem is to find out the flock patterns without any duplication or overlapping. Also the number of disks produced per time interval should be reduced so as to reduce cost of joining these disks for finding patterns between time intervals. To overcome duplication and overlapping, clustering methods needs to be implemented and the clustering algorithm used must be efficient and able to handle large databases. If clustering is used, then the number of flocks per time interval will be reduced and automatically, the cost of finding frequent patterns is reduced. The clustering algorithm will also ensure that no data should be contained in more than one cluster.

4. Proposed method: finding moving flock patterns using SPARROW algorithm

In the proposed approach, the spatio-temporal data is assumed to be given. Spatio-temporal data is those data which contains the details of position and time. Finding out moving flock patterns is done in five steps. i) Grouping the trajectory datasets into fixed time intervals. ii) Finding out flocks in each time interval using sparrow algorithm. iii) Developing a transactional version of the flocks. iv) Frequent pattern mining algorithm. v) post processing

.

4.1. Grouping trajectory datasets

The dataset used for the problem was obtained from Microsoft research website [9]. The dataset was collected during geolife project by 177 users from the period April 2007 to October 2011. The datasets contains information of latitude, longitude, altitude, date and time recording movement information on routine activities as well as amusement and sport activities. The dataset is available as trajectories of each user during the period. There are various other details in the dataset other than information on the latitude, longitude, altitude, date and time. We use only the details needed. Each trajectory has data in varying time periods. A snapshot of the dataset as a trajectory is given in fig 1.

For the experiment conducted in the study, the time interval is set as one month. In this example, data from

time April 2007 to March 2008 is taken. Tables for each time intervals are created.

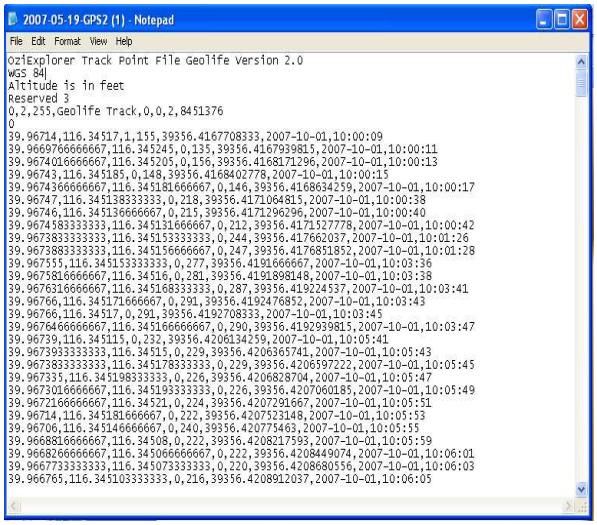

Fig 1: Sample dataset

4.2. Flock discovery

In the proposed method, flocks are literally clusters. For each time interval, clustering algorithm is applied and clusters are generated. SPAtial clusteRing algoRithm thrOugh sWarm intelligence (SPARROW) is used for this purpose [6]. SPARROW algorithm is a combination of the concepts of DBscan and adaptive flocking algorithm. SPARROW is a multi agent algorithm creating clusters in a parallel manner.

The algorithm starts with the initialization of random number of agents. The agents move around the area finding out certain properties of the points in the area. The model considers four types of agents, depending on the density of neighborhood points around them. The different agents are characterized by different colors: red agents, indicating interesting patterns of data, green, a medium one, yellow, a low one and white, indicating uninteresting zones.

The rules for determining the color of the agent are

i) Density> Minpts
   ⇨ mycolor = red (speed=0)
ii) Minpts/4<Density<Minpts
   ⇨ mycolor = green(speed=1)
iii) 0<Density<Minpts/4
   ⇨ mycolor=yellow(speed=2)
iv) Density = 0
   ⇨ mycolor=white(speed=0)

The features of alignment, cohesion and separation are taken from the flocking model and incorporated into the SPARROW algorithm. Alignment is the ability of agents to align with each other, cohesion is the ability to cohere together and separation is the ability to keep a separation distance with each other.

The main idea behind the algorithm is that red agents indicate interesting regions and white agents indicate uninteresting regions. Red and white agents stop moving where as the green and yellow agents continue to fly trying to find out denser regions.

The direction and speed of the moving agents is calculated each time. The speed of the red and white agents is given value zero, green agents given the speed 1 and yellow agents with speed 2. The direction of agents is calculated using the concepts of alignment, cohesion and separation.

The new position of the agents is given as

$$\forall i = 1,\ldots..,d \;\; x_k^i(t+1) = x_k^i(t) + v * dir_k^i \qquad (1)$$

where v is the speed of the agent , x_k^i the position of the k^{th} agent in i axis, dir_k^i the direction of movement of agent k in i axis.

For each iteration, the direction is computed by summing three components of alignment, cohesion and separation.

$$dir_k^i = dir\_al_k^i - dir\_sep_k^i + dir\_co_k^i \qquad (2)$$

The direction of alignment is given as

$$dir\_al_k^i = \frac{1}{|Neigh(green,B_k)|} . \sum_{B\alpha \in Neigh(green,Bk)} dir_k^i \qquad (3)$$

where Neigh(*col,*x) function finds out the neighbor points of the agent x with the specified color whereas Neigh(x) finds the all the neighbor points of x.

The equation for the direction of cohesion is

$$dir\_co_k^i = dir(centr(green,B_k),B_k)^i + attr\_red - rep\_white \qquad (4)$$

$centr(green,B_k)$ is the centroid of the green agents within the neighborhood.

Centroid is given as

$$\frac{1}{|Neigh(green,B_k)|} . \sum_{B\alpha \in Neigh(green,Bk)} x_k^i$$

attr_red is given as

$$\sum_{B\alpha \in Neigh(red,Bk)} dir(B_\alpha,B_k)^i$$

rep_white is given as

$$\sum_{B\alpha \in Neigh(white,Bk)} dir(B_\alpha,B_k)^i$$

The direction of separation is calculated as

$$dir\_sep_k^i = \sum_{B\alpha \in Neigh(Bk), dist(B\alpha,Bk)<dist\_min} dir(B_\alpha, B_k)^i \quad (5)$$

dir(x,y) denotes the euclidean distance between x and y.

The overall pseudo-code of the algorithm is presented below.

for i=1…. Maxiterations

1. **for each** yellow and green agents
   increase the age
   if(age>Max_life)
   kill the agent and generate a new agent in random position
   endif
   find out the color of new agents
   end foreach

2. **for each** yellow,green agents
   compute directions using the rules given
   end for each

3. **repeat for all agents**
   for all yellow and green agents, move agent in the calculated direction and with corresponding speed
   for all white agents, kill the white agent and generate a new agent in random position
   for all red agents, kill the red agents and generate a new agent close to it
   end repeat

end for (main program)

In the experiment done, after the datasets had been grouped into specific time intervals of one month, SPARROW was applied to the dataset in each time interval. Fig 2 and fig 3 depict the clusters formed in the time intervals May 2007 and August 2007

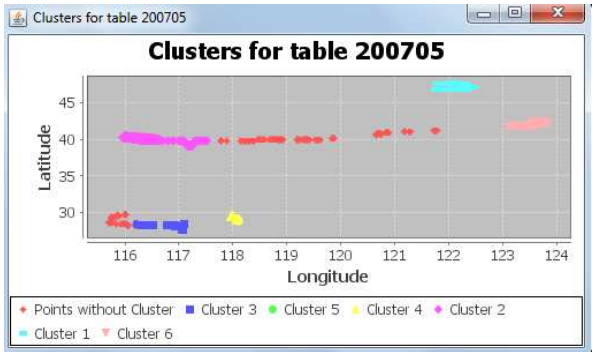

Fig 2: Clusters for the month May,2007

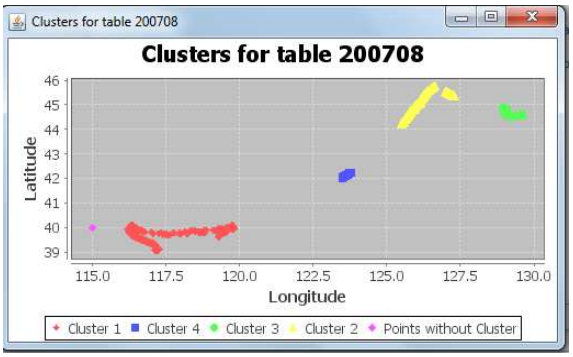

Fig 3: Clusters for the month August, 2007

4.3. Transactional version of the dataset

After the implementation of the two processes, the result will be a set of clusters for each time interval. The clusters through which each user passes are found out as the next step. For this, we go through all the clusters so far found for any element for each user.

In the experiment done, the transactional version of the trajectory dataset is obtained by checking all clusters each user passes through. From the results obtained, it was observed that some users do not pass through any clusters. Table 1 represents a portion of the transactional version of the dataset. Only the clusters of a portion of users who are similar are shown. The clusters include those from all time intervals.

Userid	Clusters
51	C1,C2,C3,C4, C5
55	C6,C7
57	C6,C7
99	C1,C2,C3,C8,C4
172	C1,C2

Table 1: Transactional version of the dataset

4.4. Frequent pattern mining algorithms

Various researches have been made on the discovery of interesting patterns. FP-growth algorithm is one of the popular methods of frequent item set mining [10]. FP-tree is a prefix tree representation of the transactional database. User specified percentage values are given for support and confidence fields. The threshold value for frequency is calculated as support/100*total number of transactions in the database. Item sets of frequency lesser than the threshold frequency are removed from the transactions and the transactions are rewritten. The support is taken

as 30% in this example and the minimum support frequency is 2. The transactional version of the dataset after removing the infrequent items C5 and C8 and after ordering the data in the decreasing order of frequency is shown in table 2. A header table is created with an entry for each item set and a link to the first item in the tree with the same name.

Userid	Clusters
51	C1,C2,C3,C4
55	C6,C7
57	C6,C7
99	C1,C2,C3,C4
172	C1,C2

Table 2: Reduced transactional database

The next step is to create the FP-tree. FP-tree is basically a prefix tree. Each path in the FP-tree indicates a set of transactions that share the same prefix and each node indicates on item. The link is from child to parent. The transactions are represented in the FP-tree. Each node carries the item set name and the frequent till then. For each transaction, it checks the tree for prefixes. If a prefix path is found out, the rest of the transaction is added to the FP-tree as a continuation of the prefix path. Otherwise, a new path is added to the tree.

The header table and the FP-tree corresponding to the data in table 2 are shown in fig 4.

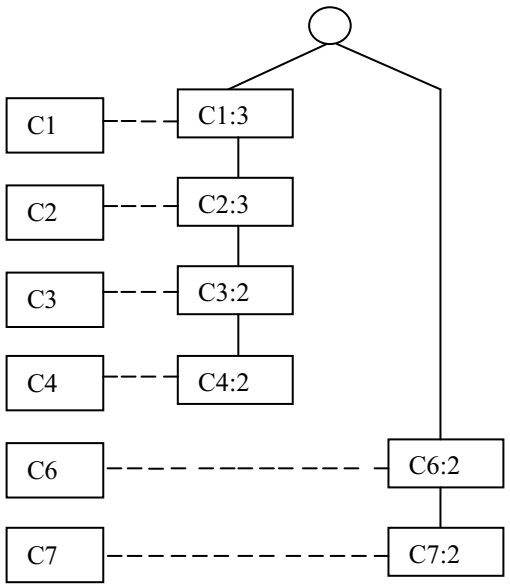

Fig 4: FP-tree for reduced transaction database

The next step is finding out frequent patterns. For this, we consider paths from every node to the root and the frequency of the patterns are also recorded. Infrequent item sets, the item sets with frequency less than the threshold frequency, C5 and C8, are removed.

4.5. Post processing

The set of frequent patterns of trajectories is checked and the consecutiveness of the patterns is verified. Along with the cluster identification, the start and end times of the time interval are also returned. The patterns are checked so that the end time of one cluster in the pattern is consecutive to the start time of the next cluster in the pattern. This ensures valid pattern formation.

5. Implications and possible applications

The flock movement could bring into light trends and interaction among the people. By using people's movement history, we can find similarity in the movement of different users and can give personal recommendations to those similar users. These movement patterns can help decision makers involved in urban planning and transport systems. Also, this knowledge can be used in the fields of prediction and forecasting. Prediction of destination of a driver is such an example [8]. Knowing the flock patterns is an initial step for the discovery of trends in movement of the people. We can also check for any correlation between places in the patterns which has various applications as in location-based services and ubiquitous computing.

6. Experimental results

The clusters obtained were named according to their date and month. The clusters are added to the database along with the elements in it, the time interval and the user to which each data belong.

The FP-tree growth algorithm is applied to all the 12 time intervals. From the given dataset from April 2007 to April 2008, 3 flock movement patterns were obtained. The visualization of the patterns are shown in fig 5.

The first pattern is a set of the clusters in the months April 2007, May 2007, June 2007 and July 2007. The second patterns contain the clusters of July, 2007 and August, 2007. The third pattern consists of clusters of November 2007, December 2007, January 2008 and February 2008.

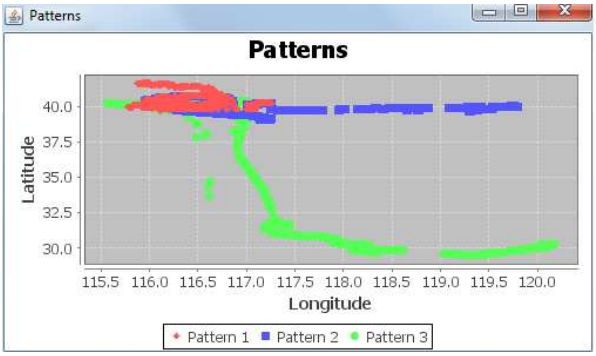

Fig 5 : Patterns

6. Conclusion

This paper explains an effective method of finding moving flock patterns using clustering algorithms. The basic steps of the proposed system are explained in detail. The datasets was divided into 12 time intervals, SPARROW algorithm was applied to each time interval, transactional version of the dataset was created and frequent pattern mining was applied to the transactional version of datasets.

The experimental results showed that the SPARROW algorithm could be applied to very large datasets. Also, it was ensured that one point would only be present in one cluster which reduced the problem of overlapping and duplicate patterns.

The main disadvantage with the SPARROW algorithm is that the parameters needed to be specified by the user for each time interval. The same parameters may not find clusters for all time intervals. As a modification, a version of the algorithm finding out the parameters automatically can be proposed further. Also, other pattern mining algorithms can be used and checked for the performance.

7. References

[1] H.Dettky, G.Ericson, L.Edenius," Real time moose tracking: an internet based mapping application using GPS/GSM collars in Sweden", *Alces* 40 (2004), pp 13-21.

[2] M.R. Vieira, P. Bakalov, and V.J. Tsotras. "On-line discovery of flock patterns in spatio-temporal data". *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM. 2009, pp. 286–295.

[3] Andres Oswaldo Calderon Romero, "Mining moving flock patterns using spatio-temporal datasets ", March 2011

[4] David Hand, Heikki Mannilla, Padhraic Smith, *Principles of Data Mining,* MIT Press, London, 2001.

[5] J. Han, M. Kamber, and A.K.H. Tung. "Spatial clustering methods in data mining: A survey". *Geographic Data Mining and Knowledge Discovery.* Taylor and Francis 21 (2001).

[6] Gianluigi Folino , Agostino Forestiero, Giandomenico Spezzano, "An adaptive flocking algorithm for performing approximate clustering", *Information Sciences*, Elsevier, 2009, pp 3059-3078.

[7] K.R. Arrigo,G.L. vanDijken, D.G. Ainley, M.A. Fahnestock, and T.Markus. "Ecological impact of a large Antarctic iceberg".Geophysical Research Letters 29.7 (2002), p. 1104. issn: 0094-8276.

[8] J.Krumm and E.Horvitz. "Predestination: Where do you want to go today?", Computer 40.4 (2007), pp. 105–107. issn : 0018-9162.

[9] Microsoft Research Asia. GeoLife GPS Trajectories. 2011.http://research.microsoft.com/en-us/ downloads/ b16d359d-d164-469e-9fd4-daa38f2b2e13/default.aspx

[10] C. Borgelt. "An Implementation of the FP-growth Algorithm". *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations.* ACM. 2005, p. 5.