Trajectory Pattern Mining Using Sequential Pattern Mining and K-Means for Predicting Future Location

G Kautsar[1]**, S Akbar**[1]

[1]Informatics, School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Ganesha Street No. 10 Bandung, Indonesia

Email : gifarikautsar@gmail.com, saiful@informatika.org

Abstract. Sequential pattern mining is a method used to find patterns while concerning the sequence of an item set. Sequential pattern mining can be used to find trajectory patterns in moving object data. To implement it in the real life, the spatial attribute of the data needs to be generalized/grouped. In this paper, K-Means is used to group the spatial attribute. In order to group the spatial attribute, the temporal attribute is also considered to see how the patterns are related to time. The resulting trajectory patterns are then used to visualize the habit of the moving object. Therefore, trajectory patterns are used as the reference in this paper to predict the future location of the object. Predicting the future location of the object is performed using the movement history of the object. Result of this research is trajectory pattern which repeat at certain time duration according to its data characteristics.

1. Introduction

Lately, people use cell phone to support almost all of their activities. Therefore, it is normal that the cell phone usage in this era is growing very rapidly. In the year of 2000, Indonesian mobile phone users were about 3.6 million. In 2012, that number increased significantly, reaching a total of 280 million [1]. Cell phone service provider always keep a record every time its user communicates with other users, e.g. when the user make a phone call or send a text message. That data is called Calls Details Record (CDR).

Meanwhile, Location Based Service (LBS) is a service that provides information about a place [2]. LBS can be easily accessed using mobile devices. An example of a system applying LBS is Global Positioning System (GPS). GPS can be used on a vehicle, cell phone, and even animal to find its position at a time. So, using GPS can generate data containing the location of an object. An example of the data is path traversed by a taxi, the migration of animals, and the location of a plane.

Spatio-temporal database is a database managing the spatial attribute that changes over time [3]. A moving object data is an example. CDR and GPS data are another example because the data describes the location of an object at a time. A set of locations in a period of time can show the movement of the object from time to time.

Pattern recognition is a technique to classify or to describe object based on its quantitative characteristics or its main property. Pattern that can be recognized form moving object data includes repetitive pattern, relationship pattern, and frequent trajectory pattern [4]. Repetitive pattern is a pattern formed from an object doing some movement periodically. An example is a bird that every day is flying from its nest to look for food. Relationship pattern is a trajectory pattern resulting from a collection of objects that is moving and each has a relation with each other. Frequent trajectory pattern is a pattern which is the trend of the movement of all objects.

Frequent trajectory pattern mining can be applied in both CDR and GPS data to see the habit of the moving object. This habit can be represented as trajectory pattern. The discovered trajectory patterns are then can be used to predict the future location of an object. In more advanced development, the discovered trajectory patterns can be utilized to know which road segment is often passed by. So, it can be used to give recommendation which road segment needs restoration or to give new routes recommendation. For example, from the trajectory patterns discovered, it is known that movement from location A to location B is quite crowded and the road segment is narrow. From that, the possible recommendation is to improve the road segment from location A to location B.

2. Preliminary Concept

2.1. Sequential pattern mining

Item is the basic value for numerous data mining problems. It can be considered as objects bought by customer, location visited by bus, etc. Item set is a set of items that are grouped by timestamp. Meanwhile, data sequence is a sequence of item set associated to an object [5]. On Table 1, data sequence for object *C2* is “(Camcorder, MiniDV) (DVD Rec, DVD-R) (Video Soft)”. It means that the customer bought camcorder and miniDV on the same day, then bought DVD recorder and DVD-R on the next day, and bought video software some days later.

Sequential pattern can be found in a data sequence. For example, “(MiniDV) (Video Soft)” can be found in *C2*, but “(DVD Rec) (Camcorder)” cannot be found in *C2* because it is not in the right order (Camcorder came first). In sequential pattern mining, user has to define the minimum support value. The minimum support value is the minimum number of data sequence that satisfies the sequential patterns so the sequential pattern can be categorized as a frequent pattern. For example, from Table 1, there are four customers in four days. If the value of minimum support is 50% of total customers, then sequence pattern has to be found in minimum two data sequences. A sequential pattern mining performed on the data will find three patterns:

- S1: “(Camcorder, MiniDV) (DVD Rec, DVD-R)”
- S2: “(DVD Rec, DVD-R) (Video Soft)”
- S3: “(Memory Card) (USB Key)”

Table 1. Data sequences of four customers over four days [5].

Cust.	June 04, 2004	June 05, 2004	June 06, 2004	June 07, 2004
C1	Camcorder, MiniDV	Digital Camera	MemCard	USB Key
C2	Camcorder, MiniDV	DVD Rec, DVD-R	-	Video Soft
C3	DVD Rec, DVD-R	MemCard	Video Soft	USB Key
C4	-	Camcorder, MiniDV	Laptop	DVD Rec, DVD-R

Pattern S1 can be found at data sequence C2 and C4, pattern S2 can be found at C2 and C3, and pattern S3 can be found at C1 and C2.

2.2. K-Means

Clustering or cluster analysis is the process of grouping set of objects into clusters. The objects grouped in a cluster will have a high similarity with each other and are different with objects from the other clusters [6]. The higher the similarity and the higher the differences, the better the cluster is.

Prototype-based clustering is a type of clustering. On prototype-based clustering, a cluster is a set of objects with higher similarity to its prototype than to other prototypes [7]. K-Means is an example of prototype-based clustering.

The first step of implementing K-Means is choosing the initial K centroid, with *K* is the number of cluster defined by the user. Each object is assigned to a cluster with the closest centroid. After the assigning process is done, each cluster centroid is then updated based on the objects in the cluster.

These steps are repeated until there are no object moving to other cluster, or in another word, the process will continue until the centroid value doesn't change anymore.

Example of applying K-Means algorithm in clustering object can be seen at Figure 1. On Figure 1 (a), three objects are assigned as the initial centroids. Each object aside from the three before are then assigned to the closest cluster (based on the distance to centroids). On Figure 1, objects are represented in difference shapes, which are: triangle, square, and circle. After the objects are clustered, centroids value is updated and then the objects are clustered again based on the distance to the new centroids. On Figure 1 (b), (c), and (d), it can be seen that the centroids are moving to smaller cluster. K-Means algorithm is stopped (on Figure II.1. (d)) because the objects are no longer moving.

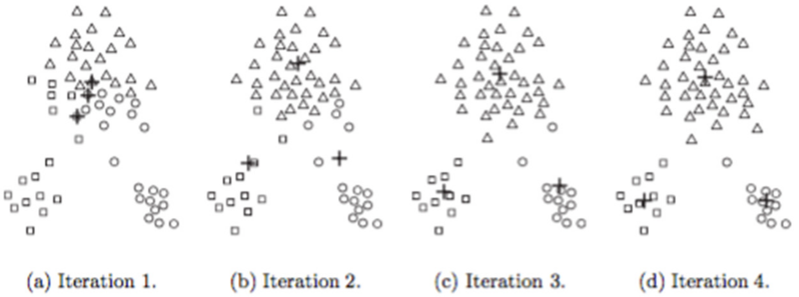

Figure 1. An example of clustering objects using K-Means.

3. Related Work

Spatial or spatio-temporal data clustering have been developed by Tork in 2012 using K-Means algorithm. Spatial data clustering method have the same concept like normal clustering. The differences are only on the data type [8]. Data shown in Figure 2 is a sample of simple data (containing no spatial data). Each point is represented in <*x,y*> form in a 2-D space. It doesn't have any information about the spatial or temporal data of the points. Using this definition, spatial clustering can be simplified into two dimensional vector, like <*x,y*>, but its values are replaced with spatial value, e.g. longitude and latitude.

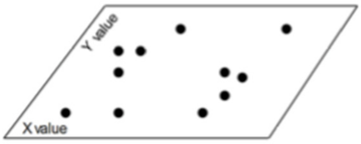

Figure 2. Simplified classical data in 2-D space [8].

Spatio-temporal data clustering is very much the same with with spatial data clustering. The differences between spatio-temporal data clustering and spatial data clustering lies in the distance calculation. For example, take a look at the K-Means distance calculation algorithm. The first step is to change spatio-temporal data into three dimensional vector <*x,y,t*>. Tuple <*x,y*> is the spatial attribute, and *t* is the temporal attribute [8]. Using the new three dimensional vector <*x,y,*t>, spatio-temporal distance (using Euclidean distance) can be calculated as:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (t_2 - t_1)^2} \quad (1)$$

Trajectory pattern mining using sequential pattern mining method has been developed by Chung et al. in 2002. The steps are as follow.

1) Make some adjustment to the data representation and arrange the database.
2) Generalize the spatial data.
3) Moving sequence/trajectory extraction. On this step, user have to define the maximum duration between each movement.
4) Trajectory pattern mining from moving sequence/trajectory data. F_k represents pattern-k obtained from data. C_k represents pattern candidate. C_k is self-join from F_{k-1}, ..., F_{k-1} * F_{k-1}. The

support value is calculated by checking if the trajectory data satisfies the pattern candidate. If the support value is less than the minimum value, pattern candidate is not assigned as frequent trajectory pattern.

4. Trajectory Pattern Mining

Trajectory pattern mining consists of several steps which are: data pre-processing, clustering the spatial attribute, data adjustment, and pattern mining. On the pattern mining step, a model of the spatial attribute cluster and trajectory pattern are gained. The result is used as reference to predict the future location.

4.1. Data preprocessing

Data preprocessing is a must before the trajectory pattern mining is performed. It consists of two steps. The first step is to cleanse the data and the second step is to generate additional attribute from temporal attribute. The cleansing is done by changing the time attribute into datetime attribute. Datetime is an attribute formed by joining the time attribute and the date attribute. Another adjustment is done to the attribute at each record to fit its definition. The last adjustment is handling the incomplete record or empty record. In the incomplete record case, if the missing attribute is the spatial or the temporal attribute, then the record won’t be used in the calculation. This is because trajectory pattern mining can’t be done without it.

For clustering step, we need an additional attribute which represents the temporal attribute. The temporal attribute is in datetime type. As it is, temporal attribute is unique and it would be harder to find the pattern. For clustering step, temporal attribute is derived to get a new attribute. A new derived temporal attribute is needed in order to enhance the quality of the trajectory pattern. To simplify the pattern mining, the temporal attribute is changed into two types of hour unit. The first type, called the *daily* type, only uses the hour attribute from the original datetime. The second type, the *weekly* type uses the date and the hour attribute from the datetime. The steps to change the temporal value into *weekly* type are as follow.

1) Change the date into day. For example, for June 14th 2016 is Tuesday.
2) Change the day value into *dayOfWeek* (see Table 2).
3) The new temporal value, the weekly type is calculated with the formula (2) below.

$$t_{weekly} = (dayOfWeek * 24) + hour \tag{2}$$

dayOfWeek is multiplied by 24 because there are 24 hours in a day.

Table 2. Day to dayOfWeek conversion.

Day	*dayOfWeek*
Monday	0
Sunday	1
Tuesday	2
Wednesday	3
Thursday	4
Friday	5
Saturday	6

4.2. Spatial clustering

There are two experiments performed to cluster the data. In the first experiment, data clustering only involves spatial attribute. K-Means algorithm is chosen because it uses the same mechanism to cluster spatial data only and spatial with temporal attribute. In the first experiment, the spatial attribute has latitude and longitude values. Formula (3) is used as distance calculation function in K-Means for spatial data clustering.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3}$$

In the second experiment, the data clustered are spatial attribute (latitude and longitude) and temporal attribute (which value had been changed into daily or weekly). Formula (1) is used as distance calculation function for the second K-Means experiment. Spatial attribute (x,y) and temporal attribute (t) each has different unit, but using distance calculation function (see Formula 1), both attributes are considered as equal. Normalization of attribute is done to balance the spatial and temporal attributes. With normalization, there are no specific attribute that would affect the cluster arrangement greatly. The result from data clustering step is a data clustering model. This model is used to define test data cluster before it is used later for predicting future location.

4.3. Generating trajectory data

The trajectory data is generated from clustered data. Object id, datetime, and spatial cluster are the attributes used in this step. Spatial attribute is no longer used because it has been replaced by the spatial cluster. In detail, trajectory data is generated from set of spatial clusters with the same object id in a certain time range. Set of trajectory data can be generated using these following steps.

1) Define the maximum movement duration from one location to the next location.
2) Calculate time differences between location *L* and last location in trajectory *T*.
3) If the time difference is less than or equal with maximum movement duration, then *L* is the next location of trajectory *T*.
4) If the time difference is more than the maximum movement duration, then *L* is the first location of a new trajectory.

4.4. Trajectory pattern mining

Temporal pattern mining, which was developed by Chung et al., 2002, is used for trajectory pattern mining. Temporal pattern mining is sequential pattern mining for moving object data. Temporal pattern mining is chosen because trajectory pattern can be recognized even though there is a missing point in the data. Data used in pattern mining are the data that have been generated into trajectory data. Some examples of generated trajectory data are <A B C D>, <A E C D>, and <B C D A>. The result of pattern mining using temporal pattern mining for the example trajectory data above can be seen at Table 3.

Table 3. Examples of trajectory pattern obtained from trajectory data.

No.	Pattern	Obtained from trajectory...
1.	<A C>	<**A** B **C** D>, <**A** E **C** D>
2.	<A D>	<**A** B C **D**>, <**A** E C **D**>
3.	<C D>	<A B **C D**>, <A E **C D**>, <B **C D** A>
4.	<B C>	<A **B C** D>, <**B C** D A>
5.	<A C D>	<**A** B **C D**>, <**A** E **C D**>
6.	<B C D>	<A **B C D**>. <**B C D** A>

5. Location Prediction

Location prediction is initiated with data pre-processing. Data pre-processing is performed by clustering spatial attribute using clustering model produced in trajectory pattern mining step. After the data is clustered, trajectory data is generated using the same steps used in trajectory pattern mining. Location prediction is done with the following steps.

1) Each trajectory pattern is divided into two parts, the first part is called *route* and the second part is called *nextStep*. The *route* is a set of spatial cluster from the first cluster to *n-1* spatial cluster from the trajectory pattern. *nextStep* is the last spatial cluster from trajectory pattern. For example, for trajectory pattern <S_1 S_2 S_3 S_4 … S_{n-1} S_n>, the value of *route* is <S_1 S_2 S_3 S_4 … S_{n-1}> and the value of *nextStep* is <S_n>.
2) Testing is performed to measure the similarity between test data and *route*. If *route* is a sequence pattern from test data, the prediction of future location is the *nextStep* of its *route*.
3) Similarity value of test data and *route* can be measured by two type of pattern quality functions, *value* and *density*. *Value* type is the value of support from trajectory pattern, while

support value is the number of trajectory data which satisfies trajectory pattern [10]. Meanwhile, value of *density* type is obtained using Formula 4.

$$density = support * \frac{\overline{route}}{trajectory\ data\ test} \tag{4}$$

4) If there is more than one route which satisfies the test data, then a route with the highest value of pattern quality will be chosen. Table 4 shows the examples of trajectory patterns.

Table 4. Example of set of trajectory pattern.

Pattern	*support*	*route*	*nextStep*
<A B A>	10	<A B>	<A>
<A D B C D>	6	<A D B C>	<D>
<B A C>	7	<B A>	<C>
<D B D>	3	<D B>	<D>

For example, <C D A D B C> is the test data to be predicted. In Table 5, examples of the similarity calculation between test data and trajectory pattern from Table 4 are shown. The prediction result using *value* type of pattern quality calculation is <A>, while the prediction result using *density*-typed pattern quality is <D>.

Table 5. Examples of similarity calculation between test data and trajectory pattern.

route	*nextStep*	Test Data	*value*	*density*
<A B>	<A>	<C D **A** D **B** C>	**10**	10*2/6 = 3,33
<A D B C>	<D>	<C D **A** D **B** **C**>	6	**6*4/6 = 4**
<B A>	<C>	-	0	0
<D B>	<D>	<C D A **D** **B** C> or <C **D** A D **B** C>	3	3*2/6 = 1

6. Experimental Result

In trajectory pattern mining, the value of some variables which will determine the trajectory pattern needs to be determined. The purpose of the experiment is to determine the best combination from the value of variables which would yield the best accuracy. Experiment is done using two datasets. The first dataset is CDR data, and the second dataset is taxi GPS data. Each dataset is divided into two parts, 80% of data as train data, and 20% of data as test data. Variables which values are needed to be determined are:

1) Number of K-Means cluster.
2) Temporal attribute usage in clustering.
3) Pattern quality calculation function.
4) Minimum support value for trajectory pattern.

Experiment is done by trajectory pattern mining on train data using every possible combination of variables value. The evaluation of experiment result is done by calculating accuracy from location prediction of test data.

6.1. Experiment using CDR data

The CDR data used for experiment is CDR data from mobile phone users in Bali and its surrounding areas. The characteristics of CDR data are as follow.

1)	Number of records	: 1.740.863
2)	Number of objects	: 20.029
3)	Average locations per object	: 87
4)	Number of distinct locations	: 2.416
5)	Data duration range	: October 6th 2014 - October 12th 2014

Values from variables that are used in experiment using CDR data are as follow.

1) Number of K-Means cluster : 10, 15, 20, 25, and 30.
2) Temporal attribute usage in clustering : *weekly*, *daily*, and *no* (without temporal attribute).
3) Pattern quality calculation function : *value* and *density*.
4) Minimum support value : 15, 20, 25, 30, and 35.

Based on experimental result, the 3 best variables configuration can be seen in Table 6.

Table 6. Best trajectory pattern mining and location prediction configuration for CDR data.

No	*Using Temporal*	*nCluster*	*Min. Support*	*Pattern Quality*	Accuracy *(%)*
1	*weekly*	10	15	*value*	99,43726
2	*weekly*	10	15	*density*	99,43726
3	*weekly*	10	20	*value*	99,3308

Temporal attribute usage in clustering greatly determines accuracy result of location prediction. *Weekly* type temporal attribute usage in clustering resulted in higher accuracy than *daily* type and *no* type. Average accuracy value for *weekly* type is 98,96%, while average accuracy value for *daily* type is 50,44% and for *no* type is 55,60%. Average accuracy value for *weekly* type is higher, that means that trajectory pattern in CDR data is unique for each day in a week (repeated every week).

Average accuracy value of location prediction based on minimum support value is shown in Figure 3. The higher the minimum support value, the smaller the average of accuracy value is. This is because higher minimum support value will cause more pattern candidates did not satisfy the minimum value.

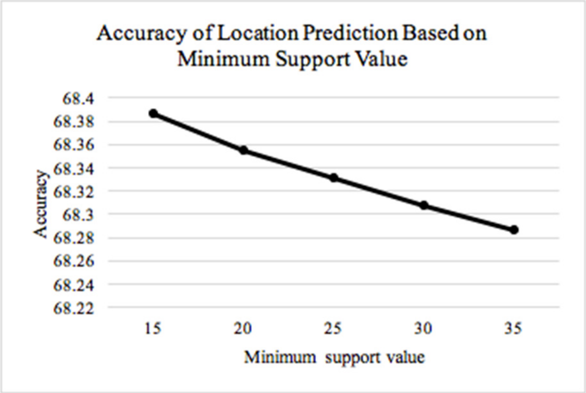

Figure 3. Accuracy of location prediction of CDR data based on minimum supprot value.

6.2. Experiment using taxi GPS data

Characteristic of taxi GPS data used in the experiment are as follows.

1) Number of object : 421
2) Number of record : 34.911
3) Average locations per record : 48
4) Number of distinct location : 1.331.454
5) Data duration range : July 1st 2013 - July 7th 2013

Taxi GPS data type is different with the general moving object data type. Taxi GPS data contains a set of spatial points which are passed in taxi trip. In other words, one record of taxi GPS data represents one trajectory, not only one spatial point. One spatial point in taxi GPS data is recorded every 15 seconds in every taxi trip.

Variables value that are used in the experiment using taxi GPS data are as follow.

1) Number of K-Means cluster : 30, 40, 50, 60, and 70.
2) Temporal attribute usage in clustering : *weekly*, *daily*, and *no* (without temporal attribute).
3) Pattern quality calculation function : *value* and *density*.
4) Minimum support value : 15, 20, 25, 30, and 35.

Based on the experimental result, the 3 best variable configurations can be seen in Table 7.

Table 7. Best trajectory pattern mining and location prediction configuration for taxi GPS data.

No	*Using Temporal*	*nCluster*	*Min. Support*	*Pattern Quality*	Accuracy (%)
1	*daily*	30	15	*value*	99,71202
2	*daily*	30	15	*density*	99,71202
3	*daily*	30	20	*value*	99,71202

Based on experimental result using taxi GPS data, temporal attribute usage in clustering really determines the accuracy of location prediction. *Daily* type temporal attribute usage in clustering has higher average accuracy value than *weekly* type and *no* type. Average accuracy for *daily* type is 99,586%, while the average accuracy value for *weekly* type is 34,296% and the average accuracy value for *no* type is 41,833%. Average accuracy value for daily type is higher, that means the trajectory pattern in taxi GPS data is repeated every day.

The effect of minimum support value in this experiment, is same as the effect of minimum support value in CDR data experiment. The higher minimum support value, the smaller average of accuracy value.

7. Conclusion and Future Works

In this research, trajectory pattern mining and future location prediction has been developed using sequential pattern mining method. Trajectory pattern mining and future location prediction has been tested using CDR data and taxi GPS data. Based on the experiment result, the best accuracy for location prediction using CDR data is 99,437% and the best accuracy using taxi GPS data is 99,712%. K-Means algorithm is used for spatial data clustering. Clustering is done twice, using only the spatial attribute and spatial and temporal attribute. Temporal attribute usage in clustering highly affects the accuracy of trajectory pattern mining. Using temporal attribute, time characteristic of trajectory pattern can be recognized. Based on experiment result of trajectory pattern mining and future location prediction, trajectory pattern of CDR data is repeated every week. Meanwhile, the trajectory pattern of taxi GPS data is repeated every day.

For future research of trajectory pattern mining, spatial clustering can be done using region/area concept. For example, spatial attribute is clustered based on administrative region.

8. References

[1] K Ariansyah 2014 Proyeksi Jumlah Pelanggan Telepon Bergerak Seluler di Indonesia *Buletin Pos Dan Telekomunikasi* **12(2)** 151-166

[2] J Schiller, A Voisard. (Eds.) 2004 *Location-based service* (Amsterdam: Elsevier)

[3] M Nanni 2002 Clustering Methods for Spatio-Temporal Data *PhD thesis, Dipartimento di Informatica, Università di Pisa*

[4] Z Li, J Han, M Ji, L A Tang, Y Yu, B Ding, J G Lee, R Kays 2011 MoveMine: Mining Moving Object Data for Discovery of Animal Movement Patterns *ACM Transactions on Intelligent Systems and Technology (TIST)* **2(4)** 37

[5] F Masseglia, M Teisseire, P Poncelet 2005 Sequential Pattern Mining *Encyclopedia of Data Warehousing and Mining* 1028-1032

[6] H Miller, J Han 2001 Spatial Clustering Methods in Data Mining *Geographic Data Mining and Knowledge Discovery* 188–217

[7] P N Tan 2006 Introduction to data mining (India: Pearson Education India)

[8] H F Tork 2012 Spatio-temporal clustering methods classification *In Doctoral Symposium on Informatics Engineering* 199-209

[9] J Chung, O Paek, J Lee, K Ryu 2002 Temporal Pattern Mining of Moving Objects for Location-Based Service *Database and Expert Systems Applications* 331-340

[10] T H N Vu, K H Ryu, N Park 2009 A method for predicting future location of mobile user for location-based services system *Computers & Industrial Engineering* **57(1)** 91-105