A Survey Paper on Trajectory Pattern Mining for Pattern Matching Query

S. R. Ghule
PG Student, Pune University

S. M. Shinde
Assosiate Prof., Pune University

ABSTRACT
Large amount information of moving objects on road network is being collected with the help of various recent technologies. The tracking of these moving objects on road networks is becoming important because of it's application in various areas.

Classification has been used for classifying various kinds of data sets like graph, text documents. However, there is a lack of study on data like trajectories on road networks. Data mining techniques, especially, sequential pattern mining can be used to extract frequent spatio-temporal patterns. Again it needs to confine the length of sequential patterns to ensure high efficiency. After extracting frequent sequential patterns in trajectories, classification can be applied to classify patterns which provide useful information in applications such as city and transportation planning, road construction, design, and maintenance, marketing sector.

In this paper, whole pattern matching query concept is adopted after the classification to find total traffic volume on given trajectory edge. At the same time, user can find number of vehicles moving in one as well as in both directions on that particular trajectory.

Keywords
Trajectory, sequential patterns, frequent pattern based classification, location based services, pattern matching.

1. INTRODUCTION
In simple words, a trajectory is a sequence of route of a moving object. Since a large amount of trajectories can be accumulated for a short period of time, many applications need to summarize the data or extract valuable knowledge from it. As a part of the trend, discovery of trajectory patterns has been paid great attention due to many applications.

1.1 Frequent patterns
Frequent patterns are itemsets that appear in a data set with frequency equal to or greater than a user-specified threshold.

1.2 Basic frequent itemset mining methodologies
Basically, Apriori algorithm is used to find frequent itemsets**.** It significantly reduces the size of candidate sets using the Apriori principle. However, it generates large number of candidate sets and repeatedly scans the database to check the candidates by pattern matching. So alternative is to use FP-growth method that mines the complete set of frequent itemsets without candidate generation. FP-growth works in a divide-and-conquer way. If the data is presented in horizontal data format, both the algorithms like Apriori and FP-growth can be used to mine frequent patterns. Similarly, algorithm like quivalenceCLASSTransformation (Eclat) can be used to extract the frequent patterns from data which is presented in vertical data format.

1.3 Mining closed frequent itemsets
A major challenge in mining frequent patterns from a large data set is the fact that such mining often generates a huge number of patterns satisfying the min_sup threshold, especially when min_sup is set low. As a result large number of patterns are generated which will contain huge number of smaller, frequent subpatterns since if a pattern is frequent, each of its sub-patterns is frequent because of low value of minimum support. To handle this problem, closed frequent pattern mining is used.

A pattern is said to be a closed frequent pattern in a data set if that pattern is frequent in data set and there will not exist any super-pattern such that it will have the same support as that of pattern found from data set.

1.4 Trajectory classification
It is defined as the process of predicting a vehicle's class label based on its trajectory. Trajectory classification can be used in various areas like traffic congestion recognition and control as well as marketing sector.

By analyzing the behavior of trajectories on road networks, the routes followed by the vehicles are critical features for classification. Also the order of these routes is important for improving classification accuracy. Based on previous study, it is concluded that (frequent) sequential patterns are good feature candidates since they preserve this order information.

Furthermore sequential pattern mining from trajectories tends to be time consuming since trajectories can be too long. So it is necessary to confine the length of trajectory. Such length-confined sequential patterns are called partial sequential patterns.

1.5 Pattern matching queries
TPM supports three types of pattern matching queries - whole, subpattern, and reverse subpattern matching for road-network trajectories. Whole pattern matching captures the global similarity between two trajectories and compares them without ignoring any part of them. Subpattern matching considers the query trajectory as whole and some interesting parts of database trajectories that best match the query trajectory and ignores rest of them. In reverse subpattern matching, the query is typically much longer than the trajectories in the database and the query retrieves the trajectories in the database that match a certain part of the query pattern.

2. RELATED WORK
Spatiotemporal data provide location as well as ordering information. In order to mine patterns from spatiotemporal data, many techniques are provided in the literature. Most of the techniques have partitioned data into disjoint cells like fixed grid. Obviously, mining patterns from each cell is much simpler than mining patterns from huge size of database.

In order to analyse large data sets of trajectories, instead of analyzing individual locations it is also necessary to aggregate

those locations into geographic regions (Giannotti et al. 2007, Lee et al. 2007, Andrienko 2010). There exist some methods to construct region from trajectory data. These existing methods normally use a density or distance based approach, in which individual locations are combined so as to form a cluster. However, such methods does not consider topological order or relations between trajectories since trajectories can be represented in graph format.

Most existing methods which are used for trajectory analysis use vector-based approaches. In this approach each trajectory is processed separately and then trajectories (or sub-trajectories) are compared and grouped together based on a vector of various characteristics like location or distance, time or difference, speed and angle

The concept of frequent pattern-based classification has been introduced in associative classification. Representative methods include CBA, CMAR, CPAR, and RCBT. In these methods, association rules are generated and analyzed for use in classification. The methods search for strong associations between frequent patterns (conjunctions of attribute-value pairs) and class labels.

Cheng et al. have proposed a frequent pattern based classification method for relational data. Efficiency issues have also been addressed in subsequent work [5]. This method uses frequent patterns as combined features. The authors have provided in-depth analysis on why frequent patterns provide a good solution for classification and investigated the relationship between discriminative power and pattern frequency. By using this relationship, the authors have developed a formal strategy for determining the minimum support threshold.

A number of frequent pattern-based classification methods have been developed in different domains. Lodhi et al. have proposed a classification method for text documents, Leslie et al. for protein sequences, and Deshpande et al. for graphs, where phrases, substrings, and subgraphs are used as features. In all these studies, frequent patterns are generated, and the data are mapped to a higher dimensional feature space. Data which are not separable in the original space become linearly separable in the mapped space. These methods, however, are not tailored to trajectory classification. Fraile and Maybank have applied a hidden Markov model (HMM) to classifying vehicles trajectories. There are more methods developed for time series and general trajectories.

The simplification or generalization of trajectories involves several different aspects. First, the route (or geometric shape) of each trajectory may be too complex or detailed and thus need simplification. For example, the algorithm like Douglas–Peucker is often used to simplify each trajectory. Simplification of trajectories involves removing points and at the same time preserve the general shape (Jeung et al. 2008). Second, even after this simplification, trajectories may still found to be too complex to compare. Therefore, trajectories can further be divided into sub-trajectories (Lee et al. 2007) and thes sub-trajectories can be further used for analysis

To measure similarities among trajectories after the simplification, one may also need to extract a vector of attributes for trajectories. For example, Dodge et al. (2009) present an approach to segment and extract local and global attributes of trajectories, such as the movement speed, duration, curvature and other descriptors. The attributes which are extracted can then be undergone through the various methods like metric similarity calculation and multivariate analysis or classification methods, such as support vector machines principal component analysis or Markov models etc.

To compare and group trajectories, either the similarity between trajectories can be defined using each trajectory as a whole or sub-trajectory attributes can be considered. For example, the partition-and-group approach partition each trajectory into number of sub-trajectories based on some geometric characteristics, and then group these sub-trajectories into clusters and finally cluster or classify trajectories based on the sub-trajectory clusters. For trajectory classification, the partition step uses class labels to improve trajectory segmentation. The clustering step used a density-based approach to group trajectories that form a dense group. There is also research on discovering patterns using different similarity measures at different cluster levels. For both of the above two steps, partitioning and grouping means simplifying individual trajectories and comparing or grouping these trajectories into clusters respectively, it is important to find regions of interest and found patterns can be generalized over the geographic space. The regions of interest can be defined subjectively by the user or derived from the data. For the latter, one option is to use density-based methods, which partition the space with predetermined grid cells, find the trajectory density in each cell and group dense cells into regions for further analysis (Giannotti et al. 2007, Lee et al. 2007, Masciari 2009). Another option is to use distance-based clustering methods, which groups points that are geographically close into clusters to simplify trajectories (Andrienko and Andrienko 2010), where one can change a distance threshold to achieve different levels of generalization.

In this section, we briefly review recent studies on trajectory pattern mining. These studies are on mining different trajectory patterns, but not on classification based on trajectory patterns. Giannotti et al. [3] have developed an algorithm of mining trajectory patterns. When generating T-patterns, similar locations should be considered as being the same since the exactly same location usually never occurs. To handle spatial similarity, regions of interest (i.e., popular regions) are discovered based on density. Then, T-patterns are represented using only these regions.

Mamoulis et al. and Cao et al.[10] have proposed methods of discovering periodic patterns in spatiotemporal sequences. Zheng et al. [8] have proposed a method to mine interesting locations and travel sequences means order of travelling route from GPS trajectories. The main idea is to exploit user's travel experiences. A user's visit to a location is regarded as a direct link from the user to the location. Then, an HITS based model is used to infer the user's travel experience and the interest of the location. However, high interestingness does not necessarily mean high discriminative power since interestingness increases as the location is visited more often.

Gidofalvi and Pedersen [4] have proposed methods of mining long, sharable patterns (LSP) from trajectories on road networks. The LSP mining algorithm tries to mine frequent patterns (not sequential patterns) mainly due to performance issues since patterns in trajectories could be extremely long. H. Cheng, X. Yan, J. Han provided solid reasons supporting frequent pattern based classification methodology. By building a connection between pattern frequency and discriminative measures such as information gain and Fisher score, they developed a strategy to set minimum support in frequent pattern mining for generating useful patterns. Based on this strategy, coupled with a proposed feature selection

algorithm, discriminative frequent patterns can be generated for building high quality classifiers.

H. Cheng, X. Yan, J. Han [5] proposed a direct discriminative pattern mining approach, DDPMine, to tackle the efficiency issue arising from the two-step approach, frequent pattern (or classification rule) mining followed by feature selection (or rule ranking).

Now a days, we can easily acquire vehicle locations from GPS. Querying over path of such trajectory is applicable in many areas. Most of the previous research focus on evaluating the spatiotemporal query such as supporting range and K nearest neighbour queries. Again many index structures are surveyed for efficient query processing on the spatiotemporal database.

Prevoius work assumes that objects can move anywhere in the space. However the movements are often constrained by obstacles. Most of the work studied moving objects with constraints on movement.

3. PROGRAMMER'S DESIGN

We present framework of frequent pattern based classification for trajectories on road network and classified patterns can be used in area like pattern matching queries.

A road network is usually modeled as a graph G(V, E), where a vertex (i.e., node) of G represents a road junction, an edge a road segment.

Features are extracted from given set of trajectories D = $\{TR_1,.....TR_{num_{tm}}\}$, with each trajectory associated with a class label ci € C = $\{C1,.....C_{num_{clag}}\}$.

3.1 Feature generation

Partial sequential patterns are generated. In order to generate partial sequential pattern from the set of trajectories, any sequential pattern mining algorithm can be used for this step.

A very important objective of this step is not to miss any sequential patterns whose discriminative power exceeds a given threshold. So, minimum support threshold can be provided to sequential pattern mining algorithms.

Another objective is to confine the length of sequential patterns. A simple heuristic approach can be used for determining the maximum length of partial sequential patterns.

3.2 Feature selection

Highly discriminative ones are selected for effective classification. The objective of the feature selection step is to select highly discriminative sequential patterns. So, F-score is used in feature selection method. The larger the F-score is, the more likely this feature is discriminative. Hence, this score is used as a feature selection criterion.

3.3 Classification model construction

Those selected patterns are fed into a classifier like SVM. The union of single features and sequential patterns is selected as features. Thus, each trajectory is mapped into a feature vector. Then feed the set of feature vectors into classifier. Each entry of a feature vector corresponds to a feature, either a single feature or a sequential pattern. The i^{th} entry of a feature vector is equal to the frequency that the i^{th} feature occurs in the trajectory.

3.4 Whole pattern matching

After classification, one can enter the trajectory on which he wants to find out total traffic volume. So, by applying whole pattern matching technique, one can find out total number of vehicles moving in one direction as well as in both direction on that particular trajectory.

4. CONCLUSION AND FUTURE ENHANCEMENT

Thus we can conclude frequent patterns provide high quality features for classification. Frequent pattern-based classification could exploit the state-of-the-art frequent pattern mining algorithms for feature generation.

Once frequent patterns are found out, these patterns can be used for various applications. So, on finding frequent path travelled by vehicle, we can publish advertisements in that area with the help of two components, one is gateway to store information of vehicles and second one is advertisement server which will send the data to gateway and finally reaches to user, may be through mobile phone.

The framework can be easily extended to use numerical attributes for classification. A set of attributes that can be attached to trajectory include the average speed, top speed, elapsed time, day, time of day etc. With the help of these attributes, the amount of traffic at specific destinations can be predicted for the near future. If we know that an area will be congested in the near future, traffic can be rerouted. Again further study on efficient and effective methods for pattern-based classification with more sophisticated patterns is an interesting direction for future research

5. REFERENCES

[1] Jae-Gil Lee, Member, IEEE, Jiawei Han, Fellow, IEEE, Xiaolei Li, and Hong Cheng, "Mining Discriminative Patterns for Classifying Trajectories on Road Networks", ieee transactions on knowledge and data engineering, vol 23, No. 5, May 2011

[2] Gook-pil Roh, Jong Roh, Hwang, "Supporting Pattern-Matching Queries over Trajectories on Road Networks", ieee transactions on knowledge and data engineering, vol 23, No.11, November 2011

[3] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory Pattern Mining," Proc. ACM SIGKDD, pp. 330-339, Aug. 2007

[4] G. Gido´ falvi and T.B. Pedersen, "Mining Long, Sharable Patterns in Trajectories of Moving Objects," GeoInformatica, vol. 13, no. 1, pp. 27-55,

[5] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative Frequent Pattern Analysis for Effective Classification," Proc. 23rd Int'l Conf. Data Eng., pp. 716-725, Apr. 2007

[6] J. Han and M. Kamber, Data Mining: Concepts and Techniques, second ed. Morgan Kaufmann, 2006.

[7] J.Chang, H.Lee, "New travel time prediction algorithms for intelligent transportation systems", Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology Volume 21 Issue 1, 2, April 2010.

[8] Y.Zheng, L. Zhang, "Mining Interesting Locations and Travel Sequences from GPS Trajectories, "Proc. 18th Int'l Conf. World Wide Web, pp. 791-800, Apr. 2009.

[9] X. Yan, J. Han, and R. Afshar, "CloSpan: Mining Closed Sequential Patterns in Large Databases," Proc. Third SIAM Int'l Conf. Data Mining, May 2003.

[10] N. Mamoulis, H. Cao, "Mining, Indexing and Querying Historical Spatiotemporal Data," Proc. ACM SIGKDD pp.236-245, Aug.2004.

[11] R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proc. 11th Int'l Conf. Data Eng., pp. 3-14, Mar. 1995.

[12] Gook-pil Roh, Seung-won-Hwang, "TPM: Supporting pattern matching queries for road-network trajectory data", EDBT 2011, March 22-24 2011, Uppsala, Sweden.