



## **A FRAMEWORK FOR DISCOVERING THE GPS TRAJECTORY PATTERNS USING FEATURE SELECTION AND APRIORI METHOD**

**V. Vetriselvi, M.Sc., M.Phil., SET.,**

Assistant Professor, Department of Computer Science, MCA and IT & Applications,  
Shrimati Indira Gandhi College, Trichy, Tamilnadu, India

**S. kiruthiga**

Research Scholar, Department of Computer Science  
Shrimati Indira Gandhi College, Trichy, Tamilnadu, India

### **ABSTRACT**

Existing prediction methods in moving objects databases cannot forecast locations accurately if the query time is far away from the current time. Even for near future prediction, most techniques assume the trajectory of an object's movements can be represented by some mathematical formulas of motion functions based on its recent movements. However, an object's movements are more complicated than what the mathematical formulas can represent. Prediction based on an object's trajectory patterns is a powerful way and has been investigated by several works. But their main interest is how to discover the patterns. In this paper, we present a novel prediction approach, namely the Hybrid Prediction Model, which estimates an object's future locations based on its pattern information as well as existing motion functions using the object's recent movements.

**KEYWORDS:** GPS Trajectory, Pattern Discovery, Feature selection, Chi-squared analysis, Apriori Association rule mining

### **1. INTRODUCTION**

Trajectory mining is an interesting data mining problem that has been studied in the context of smart cities and the Internet of Things (IoT) [1] [2] [3]. Smart cities and the IoT are indeed the way to the future as trillions of IoT devices, ranging from coffee machines to mobile objects which may or may not be inter-connected, generate enormous



amounts of data which need to be modeled and processed effectively to improve daily life [4]. For example, to optimize the commuting time to work, many sources of information including the intended route, calendar, city traffic, weather, etc. need to come together to determine a route which would be the most convenient and therefore, smart data collection, preparation and fast algorithms are needed which can work with the incoming data and propose solutions in real time.

One of the key issues in future smart cities is the incorporation of intelligence into the cities using mobile intelligence [5]. The primary source of mobile intelligence is the mobility data collected through the Internet of Things. The data is obtained from a variety of sources, e.g. moving individuals or devices, which are constantly providing location data, along with a time stamp, to some central repository. Once such data is processed, interesting information could be revealed [6], for instance, which areas in the city are witnessing an increase in activity [7], [8], the location of any traffic anomalies [9], which person or group of people are moving [10], what the popular stay points are [11], etc.

A trajectory is a time-ordered record of a moving object obtained at pre-defined discrete time intervals. However, the 'exact' location of a moving object during these intervals could be uncertain. A lot of research has focused on trajectory uncertainties with an aim to enhance the utility of trajectories. Probabilistic databases offer ways to model uncertainties using possible world semantics [12]. The uncertainties in the trajectories could be at the event level, which is the uncertainty associated with the location of the object, or at the trajectory level, which is the uncertainty associated with the path recorded as compared to the path taken, or others [13]. An interesting solution in this regard is to record the individual mobile object readings and then create complex events using probabilistic event extraction [14].

## **2. PRE-PROCESSING USING FILTERED FEATURE SELECTION METHOD**



The number of high-dimensional data that endures and is publically accessible on the internet has very developed in the past few years. Therefore, machine learning techniques have the challenge in dealing with the significant number of input features, which is modeling an attractive issue for researchers. To utilize machine learning techniques efficiently, preprocessing of the data is essential. Feature selection [15] is one of the most frequent and prominent methods in data preprocessing, and has become a necessary component of the machine learning process. It is also known as variable selection, attribute selection, or variable subset selection in machine learning and statistics. It is the method of removing irrelevant and detecting relevant features, noisy data or redundant. This technique speeds up data mining algorithms, enhances comprehensibility and predictive accuracy. Unrelated features are those that give no useful information, and irrelevant features provide no more information than the currently selected features. Regarding supervised inductive learning, feature selection presents a set of candidate features using one of the three approaches.

- The exact size of the subset of features that optimizes an evaluation measure.
- The smaller size of the subset that satisfies a certain restriction on evaluation measures.
- In general, the subset with the best commitment among size and evaluation measure.

In the process of feature selection, noise or redundant features in the data may be hinder in many circumstances, because they are not essential and relevant for the class concept such as microarray data analysis. When the number of samples is much less than the features, then machine learning gets particularly difficult, because the search space will be sparsely populated. Therefore, the model will not able to differentiate accurately between noise and relevant data. There are two major approaches to feature selection. The first is Individual Evaluation, and the second is Subset Evaluation. Ranking of the features is known as Individual Evaluation. In Individual Evaluation, the weight of an

individual feature is assigned according to its degree of relevance. In Subset Evaluation, candidate feature subsets are constructed using search strategy.

### 3. PROPOSED FRAMEWORK FOR THE TRAJECTORY PATTERNS

Following figure represents the proposed flowchart for the mining of the trajectory patterns using GPS Trajectory dataset. For finding the frequent itemset of patterns, in the stage of pre-processing, Information Gain feature selection technique has used and in the stage of pattern discovery, Apriori Association rule mining algorithm has utilized.

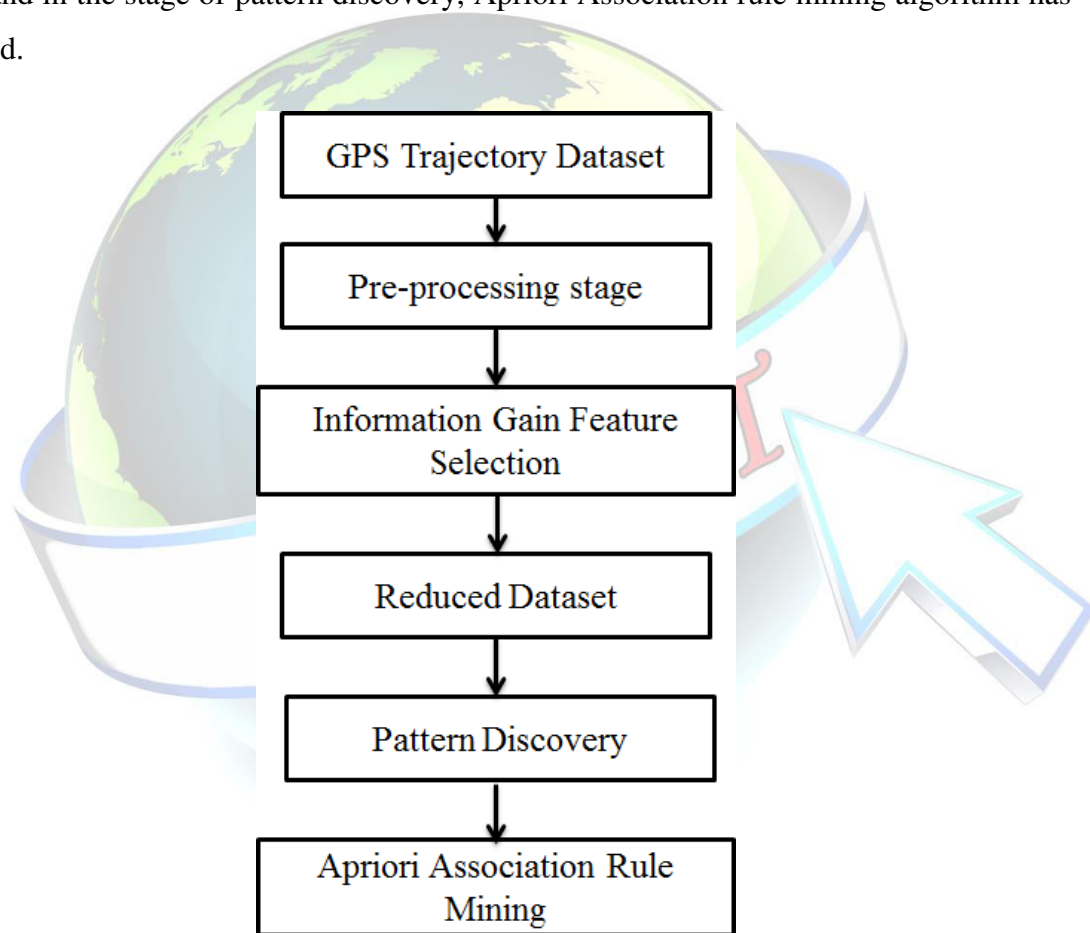


Figure 1: Proposed flowchart for the pattern discovery of GPS trajectory dataset

#### 3.1 Chi-Square Feature Selection Method



A Chi-Square Attribute Evaluation methodology is used to select the features for improving the business strategy by analyzing the GPS trajectory dataset.

Feature Selection via chi-square  $\chi^2$  test is another, very commonly used the method. Chi-squared attribute evaluation evaluates the worth of a feature by computing the value of the chi-squared statistic concerning the class. The initial hypothesis  $H_0$  is the assumption that the two features are unrelated, and it is tested by the chi-squared formula:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \left( \frac{O_{ij} - E_{ij}}{E_{ij}} \right)^2$$

Where  $O_{ij}$  is the observed frequency, and  $E_{ij}$  is the expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of  $\chi^2$ , the greater the evidence against the hypothesis  $H_0$ .

### **3.2 Association Rule Mining Via Apriori Algorithm:**

Apriori algorithm for Association Rule mining technique to produce association rule. There are different types of algorithms used to mine frequent item sets. One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. A finding frequent item set (item sets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent item-sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence.





Apriori is a seminal algorithm for finding frequent item-sets using candidate generation. It is characterized as a level wise complete search algorithm using anti-monotonicity of item-sets, “if an item-set is not frequent, any of its superset is never frequent”. By convention, Apriori assumes that items within a transaction or item-set are sorted in lexicographic order. Let the set of frequent item-sets of size  $k$  be  $F_k$  and their candidates be  $C_k$ . Apriori first scans the database and searches for frequent item-sets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent item-sets. For rule generation, precursor part is produced by utilizing apriori algorithm and for resulting; classification strategy is utilized in which the entire Telecommunication dataset is distributed into two classes, that is ordinary and attack class on the basis of marks gave in the dataset. The accompanying algorithm is utilized for finding the regular itemsets from the dataset that is Apriori algorithm:

**Algorithm:** Apriori Algorithm for finding continual itemset

**Information:** Normalize dataset, minimum support (minsupp) = 0.2

Step 1: Initialize  $k$  (number of itemset) = 1

Step 2: Find regular itemset  $L_k$  from  $C_k$  of all applicant itemsets

Step 3: Scan  $D$  and include each itemset  $C_k$ ,

Step 4: If count is greater than minimum support, then it is continuous

Step 5: Form  $C_{k+1}$  from  $L_k$ ;  $k = k + 1$



Step 6: Join Lk-1 itemset with itself to get the new contestant itemsets,

Step 7: If found a non-continuous subset then expel that subset.

Step 8: Store incessant itemset in the rule pool

Step 9: Repeat step 2 to step 9 until Ck is empty

Output: Frequent itemsets

## **4. RESULT AND DISCUSSION**

### **4.1 Description of the Dataset**

Following table 2 gives the description of the GPS trajectory dataset [R].

Table 2: Description of the GPS trajectory dataset

Sl.No	Feature Name
1	id
2	id_android
3	speed
4	time
5	distance
6	rating
7	rating_bus
8	rating_weather
9	car or bus
q	latitude
11	longitude
12	track_id
13	time

From the filtered feature selection method, the following features are selected by removing the feature with rank 0. Table 3 depicts the result obtained by chi-square feature selection method in the stage of pre-processing. Only nine features are selected among 13 features.

Table 3: Result obtained by using Chi-Square analysis feature selection method in pre-processing step

Sl.no	Original Dataset	Chi-squared feature selection
1	id	id_android
2	id_android	speed
3	speed	time
4	time	distance
5	distance	rating
6	rating	rating_bus
7	rating_bus	rating_weather
8	rating_weather	car or bus
9	car or bus	track_id
10	latitude	
11	longitude	
12	track_id	
13	time	

Following table 4a to table 4f gives the details of the Apriori Association Rule Mining algorithm for discovering trajectory patterns. Table 5 gives the trajectory pattern generated for above reduced dataset by using Chi-Square Feature Selection method.

Table 4a: Result Details of Apriori Association Rule Mining





Minimum support: 0.2 (33 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 24

Size of set of large itemsets L(3): 18

Size of set of large itemsets L(4): 7

Size of set of large itemsets L(5): 1

**Table 4b: Size of set of large itemsetsL(1): 11**

id_android=1	56
rating=2	45
rating=3	101
rating_bus=0	116
rating_bus=1	34
rating_weather=0	116
rating_weather=2	37
car_or_bus=1	87
car_or_bus=2	76
track_id=1	90
track_id=2	73



**Table 4c: Size of set of large itemsetsL(2): 22**

id_android=1 rating_bus=0 39
id_android=1 rating_weather=0 39
id_android=1 car_or_bus=2 36
id_android=1 track_id=1 38
rating=2 rating_bus=0 45
rating=2 rating_weather=0 45
rating=3 rating_bus=0 65
rating=3 rating_weather=0 65
rating=3 car_or_bus=1 57
rating=3 car_or_bus=2 44
rating=3 track_id=1 51
rating=3 track_id=2 50
rating_bus=0 rating_weather=0 116
rating_bus=0 car_or_bus=1 87
rating_bus=0 track_id=1 86
rating_bus=1 car_or_bus=2 34
rating_weather=0 car_or_bus=1 87
rating_weather=0 track_id=1 86
rating_weather=2 car_or_bus=2 37
rating_weather=2 track_id=2 36
car_or_bus=1 track_id=1 57
car_or_bus=2 track_id=2 43

**Table 4d: Size of set of large itemsetsL(3): 15**

id_android=1 rating_bus=0 rating_weather=0 39
id_android=1 rating_bus=0 track_id=1 34
id_android=1 rating_weather=0 track_id=1 34

rating=2 rating_bus=0 rating_weather=0 45
rating=3 rating_bus=0 rating_weather=0 65
rating=3 rating_bus=0 car_or_bus=1 57
rating=3 rating_bus=0 track_id=1 48
rating=3 rating_weather=0 car_or_bus=1 57
rating=3 rating_weather=0 track_id=1 48
rating=3 car_or_bus=1 track_id=1 40
rating_bus=0 rating_weather=0 car_or_bus=1 87
rating_bus=0 rating_weather=0 track_id=1 86
rating_bus=0 car_or_bus=1 track_id=1 57
rating_weather=0 car_or_bus=1 track_id=1 57
rating_weather=2 car_or_bus=2 track_id=2 36

**Table 4e: Size of set of large itemsetsL(4): 6**

id_android=1 rating_bus=0 rating_weather=0 track_id=1 34
rating=3 rating_bus=0 rating_weather=0 car_or_bus=1 57
rating=3 rating_bus=0 rating_weather=0 track_id=1 48
rating=3 rating_bus=0 car_or_bus=1 track_id=1 40
rating=3 rating_weather=0 car_or_bus=1 track_id=1 40
rating_bus=0 rating_weather=0 car_or_bus=1 track_id=1 57

**Table 4f: Size of set of large itemsetsL(5): 1**

rating=3 rating_bus=0 rating_weather=0 car_or_bus=1 track_id=1 40
---

**Table 5: Pattern Discovery of GPS trajectory using Apriori Association Rule Mining method**



Pattern of GPS Trajectory	Support and Confidence Level of Pattern
rating_weather=0 116 ==>rating_bus=0 116	<conf:(1)> lift:(1.41) lev:(0.21) [33] conv:(33.45)
rating_bus=0 116 ==>rating_weather=0 116	<conf:(1)> lift:(1.41) lev:(0.21) [33] conv:(33.45)
car_or_bus=1 87 ==>rating_bus=0 87	<conf:(1)> lift:(1.41) lev:(0.15) [25] conv:(25.09)
car_or_bus=1 87 ==>rating_weather=0 87	<conf:(1)> lift:(1.41) lev:(0.15) [25] conv:(25.09)
rating_weather=0 car_or_bus=1 87 ==>rating_bus=0 87	<conf:(1)> lift:(1.41) lev:(0.15) [25] conv:(25.09)
rating_bus=0 car_or_bus=1 87 ==>rating_weather=0 87	<conf:(1)> lift:(1.41) lev:(0.15) [25] conv:(25.09)
car_or_bus=1 87 ==>rating_bus=0 rating_weather=0 87	<conf:(1)> lift:(1.41) lev:(0.15) [25] conv:(25.09)
rating_weather=0 track_id=1 86 ==>rating_bus=0 86q	<conf:(1)> lift:(1.41) lev:(0.15) [24] conv:(24.8)
rating_bus=0 track_id=1 86 ==>rating_weather=0 86	<conf:(1)> lift:(1.41) lev:(0.15) [24] conv:(24.8)
rating=3 rating_weather=0 65 ==>rating_bus=0 65	<conf:(1)> lift:(1.41) lev:(0.11) [18] conv:(18.74)
rating=3 rating_bus=0 65 ==>rating_weather=0 65	<conf:(1)> lift:(1.41) lev:(0.11) [18] conv:(18.74)
rating=3 car_or_bus=1 57 ==>rating_bus=0 57	<conf:(1)> lift:(1.41) lev:(0.1) [16] conv:(16.44)
rating=3 car_or_bus=1 57 ==>rating_weather=0 57	<conf:(1)> lift:(1.41) lev:(0.1) [16] conv:(16.44)
car_or_bus=1 track_id=1 57 ==>rating_bus=0 57	<conf:(1)> lift:(1.41) lev:(0.1) [16] conv:(16.44)
car_or_bus=1 track_id=1 57 ==>rating_weather=0 57	<conf:(1)> lift:(1.41) lev:(0.1) [16] conv:(16.44)
rating=3 rating_weather=0 car_or_bus=1 57 ==>rating_bus=0 57	<conf:(1)> lift:(1.41) lev:(0.1) [16] conv:(16.44)
rating=3 rating_bus=0 car_or_bus=1 57 ==>rating_weather=0 57	<conf:(1)> lift:(1.41) lev:(0.1) [16] conv:(16.44)
rating=3 car_or_bus=1 57 ==>rating_bus=0 rating_weather=0 57	<conf:(1)> lift:(1.41) lev:(0.1) [16] conv:(16.44)
rating_weather=0 car_or_bus=1 track_id=1 57 ==>rating_bus=0 57	<conf:(1)> lift:(1.41) lev:(0.1) [16] conv:(16.44)
rating_bus=0 car_or_bus=1 track_id=1 57 ==>rating_weather=0 57	<conf:(1)> lift:(1.41) lev:(0.1) [16] conv:(16.44)
car_or_bus=1 track_id=1 57	<conf:(1)> lift:(1.41) lev:(0.1) [16]





==>rating_bus=0 rating_weather=0 57	conv:(16.44)
rating=3 rating_weather=0 track_id=1 48 ==>rating_bus=0 48	<conf:(1)> lift:(1.41) lev:(0.08) [13] conv:(13.84)
rating=3 rating_bus=0 track_id=1 48 ==>rating_weather=0 48	<conf:(1)> lift:(1.41) lev:(0.08) [13] conv:(13.84)
rating=2 45 ==>rating_bus=0 45	<conf:(1)> lift:(1.41) lev:(0.08) [12] conv:(12.98)
rating=2 45 ==>rating_weather=0 45	<conf:(1)> lift:(1.41) lev:(0.08) [12] conv:(12.98)
rating=2 rating_weather=0 45 ==>rating_bus=0 45	<conf:(1)> lift:(1.41) lev:(0.08) [12] conv:(12.98)
rating=2 rating_bus=0 45 ==>rating_weather=0 45	<conf:(1)> lift:(1.41) lev:(0.08) [12] conv:(12.98)
rating=2 45 ==>rating_bus=0 rating_weather=0 45	<conf:(1)> lift:(1.41) lev:(0.08) [12] conv:(12.98)
rating=3 car_or_bus=1 track_id=1 40 ==>rating_bus=0 40	<conf:(1)> lift:(1.41) lev:(0.07) [11] conv:(11.53)
rating=3 car_or_bus=1 track_id=1 40 ==>rating_weather=0 40	<conf:(1)> lift:(1.41) lev:(0.07) [11] conv:(11.53)
rating=3 rating_weather=0 car_or_bus=1 track_id=1 40 ==>rating_bus=0 40	<conf:(1)> lift:(1.41) lev:(0.07) [11] conv:(11.53)
rating=3 rating_bus=0 car_or_bus=1 track_id=1 40 ==>rating_weather=0 40	<conf:(1)> lift:(1.41) lev:(0.07) [11] conv:(11.53)
rating=3 car_or_bus=1 track_id=1 40 ==>rating_bus=0 rating_weather=0 40	<conf:(1)> lift:(1.41) lev:(0.07) [11] conv:(11.53)
id_android=1 rating_weather=0 39 ==>rating_bus=0 39	<conf:(1)> lift:(1.41) lev:(0.07) [11] conv:(11.25)
id_android=1 rating_bus=0 39 ==>rating_weather=0 39	<conf:(1)> lift:(1.41) lev:(0.07) [11] conv:(11.25)
rating_weather=2 37 ==>car_or_bus=2 37	<conf:(1)> lift:(2.14) lev:(0.12) [19] conv:(19.75)
rating_weather=2 track_id=2 36 ==>car_or_bus=2 36	<conf:(1)> lift:(2.14) lev:(0.12) [19] conv:(19.21)
rating_bus=1 34 ==>car_or_bus=2 34	<conf:(1)> lift:(2.14) lev:(0.11) [18] conv:(18.15)
id_android=1 rating_weather=0 track_id=1 34 ==>rating_bus=0 34	<conf:(1)> lift:(1.41) lev:(0.06) [9] conv:(9.8)
id_android=1 rating_bus=0 track_id=1 34 ==>rating_weather=0 34	<conf:(1)> lift:(1.41) lev:(0.06) [9] conv:(9.8)





## 5. CONCLUSION

In this paper, we presented a novel framework which forecasted an object's future locations in a manner utilizing not only motion function but also objects' movement patterns. Specifically, trajectory patterns of objects were defined and discovered. In this proposed framework, Chi-Squared analysis was used to reduce the size of the feature space. Apriori Association Rule Mining algorithm was utilized to discover the moving pattern of the moving objects. Using this pattern, the prediction of the moving objects can be analyzed.

## REFERENCES

- [1] Vera-Baquero, Alejandro, and Ricardo Colomo-Palacios. "Big-Data Analysis of Process Performance: A Case Study of Smart Cities." *Big Data in Engineering Applications*. Springer, Singapore, 2018. 41-63.
- [2] Montori, Federico, Luca Bedogni, and Luciano Bononi. "A collaborative internet of things architecture for smart cities and environmental monitoring." *IEEE Internet of Things Journal* 5.2 (2018): 592-605.
- [3] Zhang, Ningyu, et al. "Semantic framework of internet of things for smart cities: case studies." *Sensors* 16.9 (2016): 1501.
- [4] Bifulco, Francesco, et al. "ICT and sustainability in smart cities management." *International Journal of Public Sector Management* 29.2 (2016): 132-147.
- [5] Arasteh, H., et al. "Iot-based smart cities: a survey." *Environment and Electrical Engineering (EEEIC), 2016 IEEE 16th International Conference on*. IEEE, 2016.
- [6] Duan, Zongtao, et al. "Personalized service recommendations for travel using trajectory pattern discovery." *International Journal of Distributed Sensor Networks* 14.3 (2018): 1550147718767845.



[7] Cai, Haoshu, et al. "Mining frequent trajectory patterns of WIP in Internet of Things-based spatial-temporal database." *International Journal of Computer Integrated Manufacturing* 30.12 (2017): 1253-1271.

[8] Radhakrishna, Vangipuram, P. V. Kumar, and V. Janaki. "DESIGN AND ANALYSIS OF SIMILARITY MEASURE FOR DISCOVERING SIMILARITY PROFILED TEMPORAL ASSOCIATION PATTERNS." *IADIS International Journal on Computer Science & Information Systems* 12.1 (2017).

[9] Cai, Haoshu, Yu Guo, and Kun Lu. "A location prediction method for work-in-process based on frequent trajectory patterns." *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* (2017): 0954405417708222.

[10] Zhang, Rui, et al. "Understanding Trajectory Data Based on Heterogeneous Information Network Using Visual Analytics." *International Conference on Mobile Ad-Hoc and Sensor Networks*. Springer, Singapore, 2017.

[11] Khoroshevsky, Faina, and Boaz Lerner. "Human Mobility-Pattern Discovery and Next-Place Prediction from GPS Data." *IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction*. Springer, Cham, 2016.

[12] Gao, Yifeng, et al. "TrajViz: A Tool for Visualizing Patterns and Anomalies in Trajectory." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2017.

[13] Malinovskiy, Yegor, and Yinhai Wang. "Travel pattern discovery using mobile device sensors." U.S. Patent No. 9,374,661. 21 Jun. 2016.

[14] Moosavi, Sobhan, Rajiv Ramnath, and Arnab Nandi. "Discovery of driving patterns by trajectory segmentation." *Proceedings of the 3rd ACM SIGSPATIAL PhD Symposium*. ACM, 2016.



[15] Semwal, Vijay Bhaskar, et al. "An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification." *Multimedia tools and applications* 76.22 (2017): 24457-24475.

