Extracting Interesting Regions and Trips from Taxi Trajectory Data

Ammar M. Huneiti[1] & Omar Y. Adwan[1]

[1] Department of Computer Information Systems, The University of Jordan, Amman, Jordan

Correspondence: Ammar M. Huneiti, Department of Computer Information Systems, The University of Jordan, Amman, Jordan. E-mail: a.huneiti@ju.edu.jo

Abstract

The increasing availability of cutting-edge location-acquisition technologies such as GPS devices, has led to the generation of huge datasets of spatial trajectories. These trajectories store important information regarding the movement of people, vehicles, robots, animals, users of social networks, etc. Many research initiatives have applied data mining techniques in order to extract useful knowledge from this data. An important, and yet complicated, pre-processing step in mining patterns from trajectory data, is the identification of the Regions of Interest (RoI) that have been collectively navigated by a set of trajectories. The RoI's are being manually and subjectively pre-defined by a group of experts as popular regions, regardless of the actual behaviour of the moving objects. This research emphasizes the usefulness of applying an unsupervised machine learning technique, namely Self Organizing Map (SOM), in order to identify the RoI's associated with a trajectory dataset depending on the moving objects' behaviour. The research experiments were conducted using 180 thousand of the trajectories generated by 442 taxis running in the city of Porto, in Portugal, and they demonstrate the ability of SOM in identifying the RoI's and interesting taxi trips within the city.

1. Introduction

As defined by (Zheng, 2015), a spatial trajectory is "a trace generated by a moving object in geographical spaces usually represented by a series of chronologically ordered points". In a simpler words, a trajectory is a sequence of time-stamped locations i.e. if T is a trajectory then T = {l1,l2,..,ln} where ln=(xn,yn,tn). The location of the trajectory is represented by xn,yn at time tn. The increasing availability of cutting-edge location acquisition technologies such as GPS devices, has led to the generation of huge datasets of spatial trajectories. These trajectories store important information regarding the navigation or the movement pattern of a fleet of people, vehicles, robots, animals, users of social networks, etc. Many research initiatives have applied data mining techniques in order to extract useful knowledge from trajectory data, which is recognised as trajectory data mining (Chen et al., 2016; Zheng, 2015; Lin and Hsu, 2014; Jeung et al., 2011; Giannotti et al., 2007). Various stakeholders are interested in mining the movement of different types of objects, including humans, in order to extract useful information, and hence, knowledge for many different reasons. This extracted knowledge has many important applications in urban planning, traffic management, autonomous vehicles, Location-Based Services (LBS), animal protection, echo systems, counter terrorism, marketing, and many others.

In general, there are many different application-specific objectives for mining trajectory data. The foremost objective is the extraction of useful patterns through the identification of frequent movements over space and/or time (Giannotti et al., 2007). On the contrary, the recognition of abnormal behaviour by detecting outlier trajectories that have less frequent, or even, rare movements could be the objective (Kong et al. 2018). Another application may require the identification of the Regions of Interest (RoI) over geospatial or even virtual space such as the World Wide Web (Huneiti, 2012). Regardless of the objective, and in most cases, trajectory data come in huge numeric files which contain structured data. Therefore, it is not an easy task to deal with such huge and raw data sets, and many techniques were suggested in literature specifically to reduce and simplify trajectory data using, mainly, data abstraction and generalization techniques (Dewan et al., 2017).

Since taxi generated trips are very important indicator of the general geospatial and temporal movement pattern of people within a city. Mining taxi generated trajectories is a key source of information for peoples' navigation and

traffic analysis. Many recent research initiatives have specifically used taxi generated data for supporting the urban planning and development within major cities. This support can be utilized in traffic management, implementing intelligent transportation systems, enabling smart cities, social behavior analysis, marketing businesses and others.

This paper introduces a methodology for extracting interesting regions and trips from taxi trajectory data by clustering the pick-up and the drop-off GPS coordinates left by a fleet of taxi operating in the city of Porto, in Portugal. The paper is organized as follows: Section 2 is a review of the literature related to trajectory data mining and in particular taxi trajectory data and RoIs. Section 3 introduces the main approach adopted for mining interesting regions and trips from taxi trajectory data. Section 4 presents the experimental results. Finally, section 5 outlines the conclusions of this work which concludes the paper.

2. Literature Review

Trajectory data mining is concerned with extracting useful information and knowledge from trajectory data (Zheng 2015, Giannotti et al. 2007). Spatial trajectory data sets are often obtained using space and time observant sensors such as GPS devices, GSM networks, Wi-Fi receivers, etc.   As mentioned in the previous section extracting useful information from trajectory data is a cumbersome task. This is due to the huge data sets that contain the trajectories' raw data which are represented using millions of GPS locations and time stamps, which are normally stored in decimal format. In addition, these data sets also contain other supporting data in quantitative and qualitative formats.   Therefore, in order to extract useful information from these data sets, an effective pre-processing step is a fundamental requirement.

Traditional pre-processing tasks include data cleansing, reduction, transformation, normalization, and modeling. However, as mentioned in (Lee and Krumm, 2011) trajectory pre-processing is concerned with reducing the size of data required to store a trajectory in order to save storage costs and reduce redundant data. In addition, it includes filtering spatial trajectories to reduce measurement noise and to estimate higher level properties of a trajectory such as its speed and direction. More detailed trajectory pre-processing goals are outlined in (Zheng, 2015) including; noise filtering, stay point detection, compression, segmentation, and map matching.

According to (Chen et al., 2016), trajectory pattern mining methods are based on either clustering or frequency analysis of trajectories. Regardless of the method used, measuring the similarity between different trajectories is an important requirement for both methods. Although a standard trajectory similarity measurement technique does not exist, many techniques do exist that facilitate a quantitative measure of the spatial and/or temporal similarity between trajectories, including the work in (Shang et al. 2017; Toohey and Duckham 2015; Liu and Schneider 2012) and many other research initiatives.

Locations in most trajectory data sets are represented using a pair of GPS coordinates, which have a small margin of spatial displacement from the exact location due to limitations in measurement devices. Therefore, the same physical location is likely to be represented by many non-identical GPS readings. This can cause many problems especially when comparing or measuring the similarity between different trajectories. Moreover, this small measurement error is aggregated to all GPS coordinates that form a single trajectory which can be thousands of readings.

One of the techniques used to tackle this challenge is to cluster the coordinates of all trajectories in order to extract the Regions of Interest (RoI) associated with a particular trajectory data set (Reaz Uddin et al., 2011). Accordingly, every coordinate can be represented by its nearest cluster ID and trajectories can then be transformed into an ordered sequence of RoIs (Zheng et al., 2009). This method greatly simplifies the data set and produces a set of regions that can be associated with a semantically meaningful description that is connected with the real world.

Taxis are very popular mode of transport; therefore, their trajectory data contain rich information that can reflect on traffic flow, commuters' behaviour, and peoples' interest over space and time. Taxi trajectory data mining is concerned with extracting knowledge from taxi trips within a city. In general, mining taxi trajectory data aims at extracting the interesting regions, inferring trip patterns, or both (Zheng et al., 2009). Some studies incorporate their results with semantic or geographical interpretation to enrich their outcome (Yue et al., 2009). Others associate their results with a set of predefined activities (Gong et al. 2016) or even land use information (Liu et al., 2015).

Most of the coordinates associated with trajectory data have equal significance and weight. As an exception, coordinates related to taxi trajectories vary in their significance within a single trajectory. The origin and destination coordinates of a single taxi trip, which correspond to the pick-up/drop-off coordinates, respectively, have more significance than the rest of coordinates. The "in-between" coordinates represent the route of the taxi trip which is useful in applications that require the full route of the taxi trip. Therefore, mining interesting regions

from taxi trajectories involves applying specific attention to the pick-up and the drop-off coordinates (Moreira-Matias et al., 2016; Gong et al., 2016; Yue et al., 2009). Representing the taxi trajectories using their pick-up/drop-off coordinates is a very useful data generalization technique that greatly simplifies the data mining task. Subsequently, different density-based clustering techniques are applied in order to further convert neighbored or close range coordinates into clustered regions. As mentioned earlier, this is another very useful abstraction of the data (Dewan et al., 2018).

Kohonen's Self Organizing Map (SOM) (Kohonen, 1982) is a competitive Artificial Neural Network (ANN) that is classified as an unsupervised machine learning technique. Many trajectory data mining approaches such as (Ling and Delmelle, 2016; Shukla et al., 2012; Chen et al., 2008; Schreck et al., 2008) have used SOM for mining trajectory data. As far as this research is concerned, SOM was chosen as the primary clustering technique because it is an unsupervised learning technique which suits the nature of the trajectory data. It enables a spatial visualization of the clustered data as a 2-D location-sensitive grid topology which preserves the spatial autocorrelation between clusters. In order to preserve the spatial autocorrelation between clusters, SOM implements a neighborhood-based organization of clusters where data vectors can belong to a certain cluster and, although not as strong, still have an associative relationship with other vectors in neighboring clusters (Huneiti, 2012). In addition, SOM has the ability to deal with large number of clusters reaching as much as hundreds of generated clusters, which is an important advantage for clustering trajectory data. By using a large number of clusters the effect of the data uncertainty present in location-based trajectory data can be minimized (Dewan et al., 2017).

3. Methodology

The methodology introduced in this work aims at extracting interesting regions and trips from taxi trajectory data by building a pick-up/drop-off regions matrix. This matrix is built by, separately, clustering pick-up and drop-off GPS coordinates of every taxi trip in order to convert complex coordinates into more processing-friendly regions. The Self Organizing Map (SOM) artificial neural network is used to generate these clusters. Although trajectory data has a simple data structure, it is characterised by its huge size which requires an extensive pre-processing operation. Trajectory data need to be abstracted in order to reduce the amount of processed data and, consequently, enable the extraction of useful information and patterns. Moreover, the coordinates that constitute the trajectories need also to be mapped into their abstracted representation.

As depicted in Figure 1, the methodology for extracting interesting regions and taxi trips involves four main steps. First, a preprocessing step that aims at extracting the source and destination coordinates from the trajectory data set which reduces the data to a more manageable set. The output of this step are tuples of taxi trips represented as TT = <pick-up, drop-off>, where TT is a single taxi trip. At this stage, the pick-up and drop-off locations of all taxis' trips are represented using their GPS coordinates.
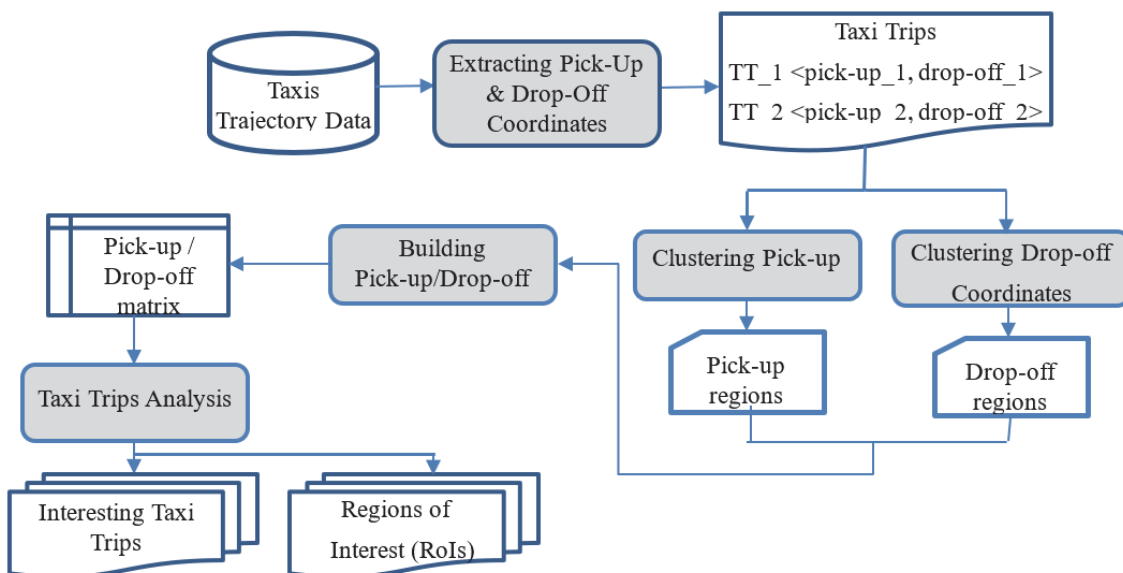

Figure 1. Methodology for Extracting Interesting Regions and Taxi Trips

Secondly, clustering the coordinates of all taxis' trips into clusters of regions with regard to the taxi's two main activities namely, pick-up and drop-off. These regions are identified by "separately" clustering the resulting pick-up and the drop-off GPS coordinates. This leads to the construction of two separate irregular grids of clusters that represent the pick-up and the drop-off regions generated by the taxis within the city. Every clustered region is given a unique ID and, consequently, all coordinates are mapped to their appropriate region (cluster) which has a unique ID. This is an abstract region based representation of the taxi trips that reduces the complexity of dealing with GPS coordinates. For all the advantages mentioned in the previous section, the Self Organizing Map (SOM), which is an unsupervised machine learning technique, was used for clustering the GPS coordinates and extracting the pick-up/drop-off regions.

Thirdly, building the pick-up/drop-off matrix where the rows and the columns of the matrix represent the pick-up regions and the drop-off regions, respectively. The integer values at the matrix cells represent the total number of taxi trips from pick-up to drop-off. This involves mapping every pick-up and drop-off GPS coordinate of every taxi trip to their region ID and incrementing the intersecting row/column value by 1. Table 1 depicts a snapshot of this matrix, where Rp_i and Rd_i are the ith pick-up and drop-off regions, respectively.

Table 1 Snapshot of pick-up/ drop-off matrix

	Rd_1	Rd_2	….	Rd_{900}
Rp_1	10	1		1220
Rp_2	0	234		5
.				
.				
.				
Rp_{400}	35	0		439

Lastly, by analysing the constructed pick-up/drop-off matrix, very useful information and patterns are extracted. These include the Regions of Interest (RoIs) and the interesting (frequent) taxi trips. This information is obtained by employing simple statistical analysis functions on rows and columns of this pick-up/drop-off matrix. In addition, the semantic interpretation of the regions is an added value to the trajectory analysis as it will be depicted in the next section.

4. Experimental Results

The experiments of this work were conducted using 180 thousand of the trajectories generated by 442 taxis owned by one of the two major taxi companies running in the city of Porto in Portugal. These were generated, approximately, in August 2013. This representative sample data set of 180 000 taxi trajectories, were selected out of the 1 710 671 trajectories available in the database which was gathered by (Moreira-Matias et al., 2016; Moreira-Matias et al., 2013).

The resulted trajectory data was put through a pre-processing stage which consisted of (i) data cleansing, getting rid of all void trajectories. For example, some trajectories have unacceptable "few" coordinates (ii) The extraction of pick-up and drop-off coordinates for every taxi ride and storing them in two separate data sets that are linked with their original trip. The pick-up and drop-off coordinates of a trajectory were extracted as the first and last coordinates, respectively, of each taxi trip.

The pick-up coordinates and the drop-off coordinates were separately clustered in order to generate the Regions of Interest (RoIs). The pick-up and drop-off RoIs correspond to the SOM's output neurons i.e. the number of SOM's output neurons reflect the number of RoIs used in each case. In most major cities around the world, as in Porto, most taxis are queued in taxi stands which are distributed around the city where they can be picked up by passengers. Other passengers may pick a taxi either by phoning the main call centre or randomly available at any street. In contrast, the drop-off locations can be anywhere in the city. Consequently, the drop-off locations are more scattered around the city than the pick-up locations which are less dispersed. Therefore, more drop-off regions were used compared to pick-up regions.

In this respect, 400 pick-up regions were used compared to 900 drop-off regions all around the city of Porto. The pick-up coordinates were used to train a 20x20 SOM output grid and the drop-off coordinates were used to train a 30x30 SOM output grid for 10 epochs each. The resulting regions were projected on Google Maps© using two separate layers. Figures 2 and 3 depict a general view of the projected pick-up and drop-off regions, respectively.

Figure 2 shows a higher density and concentration of pick-up regions around the city centre because of the existence of higher number of taxi stands. In contrast, Figure 3 shows a more dispersed drop-off regions and a more reach to the suburbs of the city which reflect the passengers' destinations.
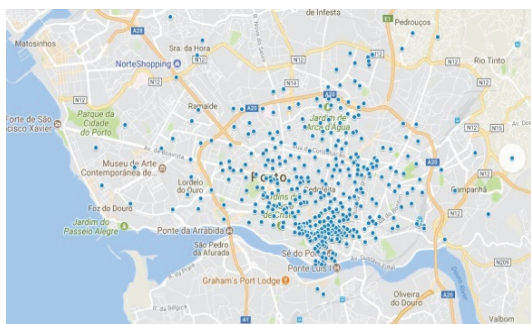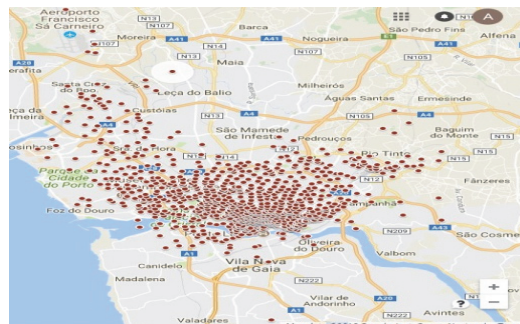

Figure 2. Projected Pick-Up Regions (400)

Figure 3. Projected Drop-Off Regions (900)

Next, the pick-up/drop-off matrix was constructed using the earlier generated pick-up and drop-off regions. A snapshot of this matrix of dimension 400x900 is represented in Table 1. The rows represent the pick-up regions and the columns represent the drop-off regions. The value of each cell represents the total number of taxi trips made from the intersected pick-up to drop-off.

This abstracted pick-up/drop-off matrix greatly simplifies the analysis of taxi trips in order to identify interesting regions and trips. For example, a descending sort of the sum of rows can identify the most popular taxi pick-up regions in the city. The top 5 pick-up regions and their semantic interpretation are shown in Table 2. The table shows the cluster/region ID, its semantic description (the name between brackets is the name of the nearest taxi stand), and the number of pick-ups initiated from this region. Visa versa, a descending sort of the sum of every column identifies the most popular taxi drop-off regions (destinations) in the city. The top 5 drop-off regions and their semantic interpretation are shown in Table 3.

Table 2. Top 5 pick-up regions

	Region (Cluster) ID	Region Description	No. of pick-ups
1	62	Campanha Train Station (Campanha)	9598
2	201	Batalha Square (Batalha)	4622
3	377	Rua de Diogo Botelho +Rua das Condominhas (Lordelo)	4087
4	285	Clerics St. + Liberty Square (Clérigos)	3806
5	18	São João Hospital (H. São João)	3755

Table 3. Top 5 drop-off regions

	Region (Cluster) ID	Region Description	No. of drop-offs
1	873	Porto Airport	7355
2	389	Campanha Train Station	4617
3	451	Empire Square (Praça do Império)	2335
4	10	Av. da República, Vila Nova de Gaia, Near El Corte Inglés de Gaia Shopping Centre	1897
5	488	Intersection Avenida da Boavista with Rua de Pedro Hispano	1768

In addition, a descending sort of the maximum of all cells' values can identify the most popular taxi trips abstracted by their pick-up and drop-off regions. The top 5 taxi trips, their semantic interpretation and the number of made trips are shown in Table 4. More taxi trips analysis actions can also be conducted on this pick-up/drop-off matrix. For example, a comprehensive analysis can be conducted for any single pick-up or drop-off region in terms of its most popular outgoing region(s) or its most popular incoming regions. In addition, non-zero average of every pick-up or drop-off region can be calculated and compared with each other. Many other statistical functions can also be

applied on this abstracted matrix as required by the application.

Table 4. Top 5 Taxi Trips (pick-up to drop-off)

	pick-up region ID	pick-up region description	drop-off region ID	drop-off region description	No. of trips
1	201	Batalha Square	873	Porto Airport	474
2	41	Freixo (Estrada da Circunvalação)	18	Escada do Codessal (Fernandina Wall)	186
3	285	Clerics St. + Liberty Square (Clérigos)	10	Av. da República, Vila Nova de Gaia, Near El Corte Inglés de Gaia Shopping Centre	174
4	62	Campanha Train Station	420	Freixo Marina	169
5	62	Campanha Train Station	369	Downtown Porto Shopping Centre (Shopping Cidade do Porto)	165

5. Conclusions

This work introduces a methodology for extracting interesting regions and trips from taxi trajectory data. The conducted taxi trajectory data abstraction was the main contribution of this work. This includes representing taxi trajectories using their pick-up and drop-off coordinates, clustering coordinates into separate pick-up regions and drop-off regions, and building the pick-up/drop-off matrix by re-joining every pick-up region with its drop-off region. This has enabled a comprehensive analysis of the taxi trips which facilitated the extraction of interesting regions and taxi trips. In addition, this research demonstrated many advantages of using SOM to cluster trajectory data. These include preserving the spatial autocorrelation between clusters and the SOMs ability to deal with large number of clusters.

Acknowledgment

References

Chen, L., Liu, L., Chen, B., & Jia, H. (2016). Research on Patterns Mining Method for Moving Objects. *Advanced Science and Technology Letters, 123*(CST 2016), 200-205. https://doi.org/10.14257/astl.2016.123.37

Chen, Y., Zhang, Y., Hu, J. (2008). Multi-Dimensional Traffic Flow Time Series Analysis with Self-Organizing Maps. *Tsinghua Science & Technology, 13*(2), 220-228. https://doi.org/10.1016/S1007-0214(08)70036-1

Dewan, P., Ganti, R., & Srivatsa, M. (2017). SOM-TC: Self-Organizing Map for Hierarchical Trajectory Clustering, Proceedings of the IEEE 37th International Conference on Distributed Computing Systems (ICDCS). https://doi.org/10.1109/ICDCS.2017.244.

Giannotti, F., Nanni, M., Pinelli, F., & Pedreschi, D. (2007). Trajectory pattern mining, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining KDD '07, Pages 330-339. http://doi.acm.org/10.1145/1281192.1281230

Gong, L., Liu, X., Wu, L., & Liu, Y. (2016). Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science, 43*(2), 103-114. https://doi.org/10.1080/15230406.2015.1014424

Huneiti, A. M. (2012). Interpreting Web Usage Patterns Generated Using a Hybrid SOM-Based Clustering Technique. *International Review on Computers and Software (I.RE.CO.S.), 7*(A)(3), 1078-1088. https://doi.org/10.13140/2.1.3359.5688.

Jeung, H., Yiu, M. L., & Jensen, C. S. (2011). Trajectory Pattern Mining. In: Zheng Y., Zhou X. (eds) Computing with Spatial Trajectories, 143-177, Springer, New York, NY. https://doi.org/10.1007/978-1-4614-1629-6_5

Kohonen, T. (1982). Self-organized formation of topological correct feature maps. *Biological Cypernetics, 43*, 59-69. https://doi.org/10.1007/BF00337288

Kong, X., Song, X., & Xia, F., et al. (2018). LoTAD: long-term traffic anomaly detection based on crowdsourced

bus trajectory data. *World Wide Web (2018) 21: 825, 21*(3), 825-847. https://doi.org/10.1007/s11280-017-0487-4.

Lee, W. C., & Krumm, J. (2011). Trajectory Preprocessing. In: Zheng Y., Zhou X. (eds) Computing with Spatial Trajectories, 3-33, Springer, New York, NY. https://doi.org/10.1007/978-1-4614-1629-6_1

Lin, M., & Hsu, W. J. (2014). Mining GPS data for mobility patterns: A survey, Pervasive and Mobile Computing. *Elsevier, 12*, 1-16. https://doi.org/10.1016/j.pmcj.2013.06.005

Ling, C., & Delmelle, E. C. (2016). Classifying multidimensional trajectories of neighbourhood change: A self-organizing map and k-means approach. *Annals of GIS Journal, 22*(3). https://doi.org/10.1080/19475683.2016.1191545

Liu, H., & Schneider, M. (2012). Similarity measurement of moving object trajectories, Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming IWGS '12, Pages 19-22. https://doi.org/10.1145/2442968.2442971

Liu, X., Gong, L., Gong, Y., & Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography, 43*, 78-90. https://doi.org/10.1016/j.jtrangeo.2015.01.016

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., & Damas, L. (2016). Time-Evolving OD Matrix Estimation using high-speed GPS data streams. *In: Expert Systems with Applications, 44*, 275-288. https://doi.org/10.1016/j.eswa.2015.08.048

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., & Damas, L. (2013). Predicting Taxi–Passenger Demand Using Streaming Data. *In: IEEE Transactions on Intelligent Transportation Systems, 14*(3), 1393-1402. https://doi.org/10.1109/TITS.2013.2262376

Reaz Uddin, M., Ravishankar, C., & Tsotras, V. J. (2011). Finding Regions of Interest from Trajectory Data, Proceedings of the IEEE 12th International Conference on Mobile Data Management, 6-9. https://doi.org/10.1109/MDM.2011.12

Schreck, T., Bernard, J., Tekusova, T., & Kohlhammer, J. (2008). Visual cluster analysis of trajectory data with interactive Kohonen Maps, Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 19-24 Oct. https://doi.org/10.1057%2Fivs.2008.29

Shang, S., Chen, L., Wei, Z., Jensen, C. S., Zheng, K., & Kalnis, P. (2017). Trajectory similarity join in spatial networks. *Proceedings of the VLDB Endowment, 10*(11), 1178-1189. https://doi.org/10.14778/3137628.3137630

Shukla, S. K., Rungta, S., & Sharma, L. K. (2012). Self-Organizing Map based Clustering Approach for Trajectory Data. *International Journal of Computer Trends and Technology, 3*(3).

Toohey, K., & Duckham, M. (2015). Trajectory similarity measures. *ACM SIGSPATIAL, 7*(1), 43-50. https://doi.org/10.1145/2782759.2782767

Yue, Y., Zhuang, Y., Li, Q., & Mao, Q. (2009). Mining Time-dependent Attractive Areas and Movement Patterns from Taxi Trajectory Data, Proceedings of the 17th International Conference on Geoinformatics, IEEE. https://doi.org/10.1109/GEOINFORMATICS.2009.5293469

Zheng, Y. (2015). Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology (TIST), 6*(3), Article No. 29. https://doi.org/10.1145/2743025

Zheng, Y., Zhang, L., Xie, X., & Ma, W. (2009). Mining Interesting Locations and Travel Sequences from GPS Trajectories, Proceedings of the 18th international conference on World wide web WWW '09, Pages 791-800. https://doi.org/10.1145/1526709.1526816

Copyrights