Study on the mining based on the improved DBSCAN algorithm in pick-up hotspots areas

Zhipu Tang[a], Xiaodong Wang[b], Hao Liu [c], Xiaowen Wang[d], Zhiqiang Wei[e]

College of Information Science and Engineering, Ocean University of China,Qingdao,Shandong,266100,China

[a]tangzhipu@ouc.edu.cn, [b]wangxiaodong@ouc.edu.cn, [c]liu.hao@ouc.edu.cn, [d]wxiaowen123@gmail.com, [e]weizhiqiang@ouc.edu.cn

Keywords:pick-up hotspots, route recommendation, DBSCAN algorithm, clustering.

Abstract. The main business model of taxies called "Roadside Beckon" adopted widely is inefficient. Passengers and drivers are ignorant of the location information of each other. The rise of various types of taxi Apps also have made an effort to solve the problem of "difficult to take a taxi". However, these methods have not fundamentally gotten rid of the taxi status of "passive waiting for passengers". Therefore, this paper proposes a taxi cruise path recommendation method based on pick-up hotspots areas. The method can recommend a cruise path with short distance and high pick-up rate for taxis. As a result, it solves the problem of "passive waiting" and aimless random roaming for guests, balance the relationship between taxis and passengers.

1. Introduction

In recent years, taxi industry in cities has developed rapidly. However, the strategy which is used to reasonably guide and dispatch taxis is delay. The main business model of taxies called "Roadside Beckon" and running mode based on sweeping street are basically "passive waiting for passengers". This phenomenon has not only caused the increase of taxi empty rate and the decrease of taxi yields, but also led to the negative effects of urban traffic [1]. Therefore, it urgently needs a strategy which is used to reasonably guide and dispatch taxis. The strategy can change the current situation of aimless random roaming for guest and balance the relationship between drivers and passengers [2].

This paper propose to make use of the improved grid DBSCAN algorithm to cluster the taxi pick-up hotspots areas. This method, which has taken the historical pick-up spot data as the data source, clusters the taxi pick-up hotspots areas and finds the core of each area [3, 4]. In order to improve the efficiency of the cruise paths recommendation, this dissertation proposes to complete the process of the recommendation with two different models, one is the off-line path construction algorithm and the other is the online recommendation algorithm [5, 6].The method can recommend a cruise path with short distance and high pick-up rate for taxis. As a result, it solves the problem of "passive waiting" and aimless random roaming for guests in inefficient ways.

2. The Improved Grid Clustering Algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm [7]. It is able to efficiently detect clusters of different sizes and shapes. However, the algorithm also has some shortcomings as follows:

(1) While clustering large amount of data, the algorithm is extremely high demands on memory and would be likely to cause the collapse of the program.

(2) When clustering uneven density or large spatial difference, DBSCAN leads to poor quality clusters.

As the dissertation is to survey and analyze multi-day taxi pick-up hotspots, the amount of data is very huge. And the location of the taxi pick-up hotspots in different areas is also very different. So Based on the above analysis of the advantages and disadvantages of DBSCAN algorithm, this paper will cluster taxi pick-up points by using the improved DBSCAN algorithm [8]. The different density regions will be clustered independently each other. For one thing, it can make up for the adverse

effects caused by the global variable value, for another it will implement their clustering process for each divided area parallel. A large amount of data is avoided loading into memory at the same time, so that the improved algorithm can reduce the high demand on memory to process data, and improve the efficiency of the clustering process.

According to the DBSCAN algorithm, we can know that the final result of clustering uneven density or large spatial difference depends heavily on global variables Eps and MinPts values. So the idea of improved grid DBSCAN algorithm is:

Firstly, mesh the regions occupied by spatial data objects, which can relatively uniform the density of a grid. The improved grid DBSCAN algorithm divide the spatial objects into rectangular group of the small step, and measure the density based on a group [9]. Thus, the number of group shouldn't be too large or too small. Generally speaking, the relationship between the number of group and spatial objects is:

$$m \approx 1.87(n-1)^{\frac{2}{3}} \quad (1)$$

Where n denotes the total number of spatial objects and m is the number of final group. We can calculate the m by the formula, then compute the corresponding steps by:

$$d = \frac{d_{\max} - d_{\min}}{m} \quad (2)$$

Where $d_{\max}$ is the maximum values and $d_{\min}$ is the minimum values of the attribute of spatial objects. d is step length. When the size of data aggregate $D = (d_1, d_2, d_3, ..., d_n)$ is designated n, we can acquire the number of groups and step length. To judge that $d_i (1 \le i \le n)$ belongs to g_i is defined by:

$$g_i = \frac{d_i - d_{\min}}{d} \quad (3)$$

Group the spatial objects according to the formula(3), count the number of each group, and draw a histogram. Fig.1 shows an example of data distribution, a rectangle denotes a group, and the height of the rectangle reflects the number of objects included in the group. That we can deduce the density of spatial objects.

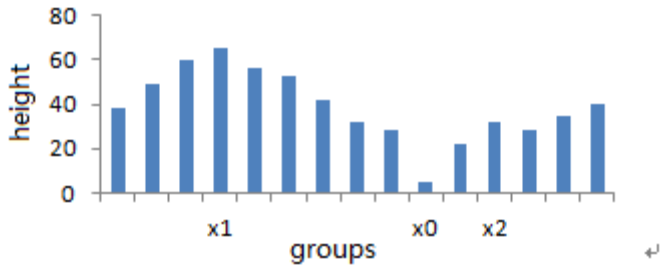

Fig.1 Data distribution

Secondly, cluster locally and analyze all of the spatial data objects in each grid based on the relative density. Divide the spatial objects into different grid by this way, then we need to cluster different grid. For large-scale uneven objects, it may result in a large number of grid area. If we cluster them respectively, the most important thing is to set the value *Eps* of each grid. The value chosen improperly may lead to low clustering quality. The RDBClustering algorithm based on relative density gets rid of the limitation of global variables. It depends only on the density of the data set. The disadvantages of RDBClustering is that it must get all of the attribute value information which clusters in the process of clustering. After gridding according to the improved DBSCAN algorithm, the number of objects in a grid will be greatly reduced. Therefore, we can do local clustering by RDBClustering algorithm.

Finally, merge the cluster in each grid based on certain conditions. Meshing may divide the objects which belong to a cluster into two adjacent grid, resulting in wrong segmentation. Therefore, we need to merge the adjacent grid as required.

3. Experiment

The experiment calculate passenger data of the corresponding time period by traversing historical data of taxis [10]. Cluster passenger point by applying the improved grid DBSCAN algorithm, and compute hot point of taxi passenger. Therefore, we need to collect and pre-process the operating data, then cluster passenger data. Experimental data is the recorded GPS data of 28,000 taxis in Beijing in May 2009. Primary data in binary text file is stored in dat format, and the GPS device upload data every 60s. Dataset includes a vehicle ID, latitude, longitude, speed, driving direction, passenger status, data acquisition time.

The area(latitude: 116.40456-116.42461, latitude: 39.89335-39.90947) adjacent to second ring, the time setting on 10 May 2009, 16:00-17:00. And select 3000 taxis that are the most number of taxi passengers, and passenger data during 16:00-17:00 in the last seven days. Then process the passenger data according to the improved grid-based DBSCAN algorithm. The gridding result of passenger data is shown in Fig.2. It is passenger data distribution of designated regions.

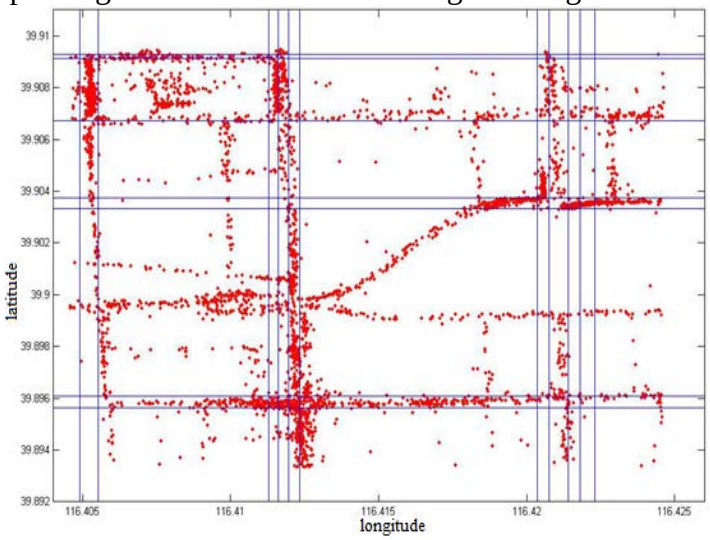

Fig.2 Passenger data distribution

Cluster passenger data by applying the improved grid DBSCAN algorithm, the hot spots of taxi pick-up passenger is shown in Fig.3.

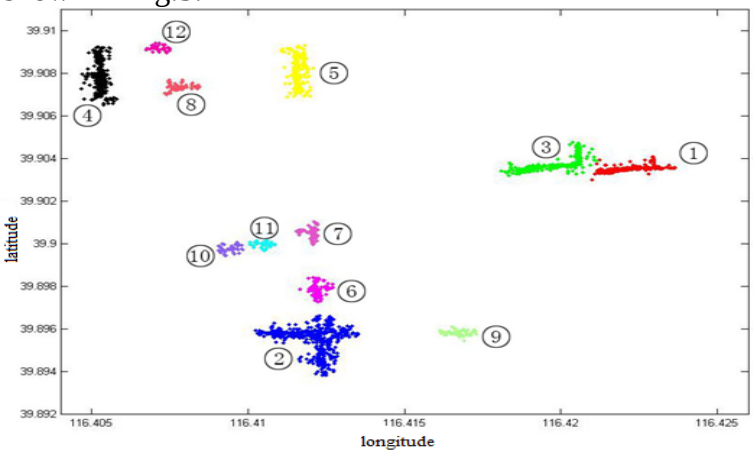

Fig.3 Passenger pick-up point

Cluster a passenger pick-up hotspots area by the improved grid DBSCAN algorithm, and calculate the core passenger point of each hotspot area. The point is regarded as the recommended point in corresponding area, and the final recommended route will be composed of this points. To efficiently compute the recommended route, we need to take into consideration the density of passenger area and the route length. The passenger points can be calculated as discussed above. All candidate for the parade route can be constructed using offline path construction algorithm. We only input the furthest distance in the client, then can calculate the best parade route successfully.

4. Conclusions

This paper combines network communication technology, GPS, GIS and spatial clustering technology to study and propose a recommendation method of calculating a taxi parade route based on pick-up hotspots areas, and it is successfully applied in the taxi system. The significance of the research is to lower taxi empty rate and shorten wait times for passengers. Although it can meet the functional requirements, it still has a long way to go.

Acknowledgements

This research is supported by National Science Foundation (61202208), and Shandong Province Science and Technology Development Plan (2014GGX101005), and Qingdao Strategic Emerging Industries Development Programme (13-4-1-45-hy).

References

[1]. Ge Y, Xiong H, Tuzhilin A, et al. An energy-efficient mobile recommender system. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. Washingtion DC, 2010, p. 899-908.

[2]. Li B, Zhang D, Sun L, et al. Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset. Pervasive Computing and Communications Workshops (PERCOM Workshops). Seattle, 2011, p. 63-68.

[3]. Yuan J, Zheng Y, Zhang L, et al. Where to find my next passenger.Proceedings of the 13th international conference on Ubiquitous computing. Beijing, 2011, p. 109-118.

[4]. Kiam T S, Nam H D, Der-Horng L. Towards An Automated Multiagent Taxi-Dispatch System. Automation Science and Engineering. Scottsdale, 2007, p. 1045-1050.

[5]. Gao H, Lim L, Wang W, et al. Pick-Up Tree Based Route Recommendation from Taxi Trajectories. The 13th International Conference on Web-Age Information Management. Harbin, 2012, p. 471-483.

[6]. Qu M, Zhu H, Liu J, et al. A cost-effective recommender system for taxi drivers. The 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, 2014, P. 45-54.

[7]. B Borah, D K Bhattacharyya. An improved sampling-based DBSCAN for large spatial databases. Intelligent Sensing and Information Processing. India, 2004, p. 92-96.

[8].Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concepts and Techniques Third Edition. Elsevier, 2011, p. 443-483.

[9]. Min Yufang, Zhao Yan, Wang Ping.GMDBSCAN: Multi-Density DBSCAN Cluster Based on Grid.e-Business Engineering. Xi'an, 2008, p. 780-783.

[10]. K D, M C. An adaptive solution to dynamic transport optimization. The fourth international joint conference on Autonomous agents and multiagent systems.New York, 2005, p. 45-51.