Research on Trip Hotspot Discovery Algorithm Based on Hierarchical Clustering

Hongqin Luo[a],Dejun Chen[b],Zhuang Xiong[c], and Kehao Wang[d]

School of Information Engineering, Wuhan University of Technology, Wuhan,430070,China

[a]1170159388@qq.com,[b]mrchendj@163.com,[c]332753291@qq.com,[d]58330573@qq.com

Keywords:travel survey, stop-point, hierarchical clustering, trip hotspot

Abstract. Aiming at the distribution characteristics of the stop-point data obtained from the mobile phone travel survey, this paper proposes an algorithm of extracting the hotspot area of urban residents by hierarchical clustering method. The algorithm is implemented and the effectiveness of the algorithm is verified by case data. The algorithm has been used in the actual travel investigation system to provide concrete decision-making basis for urban land use and transportation planning.

Introduction

The ultimate goal of the urban residents' travel investigation is to extract the statistical data of the city's macro-level traffic from the city's travel survey data, so as to understand the current traffic situation and guide the construction of the traffic.City macro-level travel information statistics are from a large number of personal daily travel data, and a traffic trip of urban residents can be characterized by a number of elements, commonly used factors include travel destination, travel start and end time, travel starting point and end, travel tool, Travel routes and travel costs.The original travel trajectory data based on the mobile travel survey is only a collection of the sensor parameters of the mobile phone, and can not express the above elements intuitively. The semantic information needs to be extracted by scientific means. Therefore, the feature extraction based on mobile travel survey has become the hotspot of the current research. Among them, the feature extraction of residents' trip hotspots is the focus of the later analysis of travel investigation. The result is of great significance to the control of urban traffic conditions.

In recent years, with the popularity of smart phones,the use of mobile phones on the GPS chip for real-time travel survey method has been rapid development.A large number of researchers through the mobile phone collection of trader and stop point data, in-depth study of its semantic features, and achieved fruitful results.Mao Haixiaodiscusses the relevant travel characteristics in the survey of urban residents' travel, and describes the influencing factors in detail,including the concept of five basic characteristics of the basic characteristics of travel, such as travel frequency, travel destination distribution, travel time distribution, travel space distribution and distribution of travel tool [1].Ramaswamy Hariharan et al. analyzed the distribution of the number of users' stop-point data of a week based on the user's all stop-point data for a week.Francesco et al. [2] proposed a method for extracting useful travel information from a large number of raw mobile data to study urban residents' travel patterns in order to use mobile data for transportation studies, the method mainly extracts the travel frequency, the travel distance and the resident position of the user, and carries on the correlation analysis based on this[3].Vangelis et al. proposed a hierarchical modular framework for extracting travel feature information from mobile phone network location data and forecasts the traffic demand in a particular area based on this[4].Hu Zhongwei et al.

proposed the technical route of the residents' travel feature extraction and analysis based on the mobile phone positioning data, and analyzed the distribution of the residents' travel OD, the population distribution and the travel characteristic parameters, and used case data for correlation analysis[5].Youngsung et al. proposed a learning model that identifies the user's travel activity type based on the user's stop-point and economic attribute [6].

How to use the stop-point data obtained by the mobile phone travel investigation to analyze the hot area of the residents' travel in the setting area is an urgent problem to be solved by the current urban planning department.Based on the hierarchical clustering algorithm, this paper proposes a trip hotspot discovery algorithm to extract the hotspot area based on the user's stop-point. This algorithm uses a large number of users' stop-point from the urban macro level to calculate the hot information of the urban residents' travel area. To find urban transport problems, and then for further follow-up urban planning to provide data support.

Trip Hotspot Discovery Algorithm Based on AGNES

City trip hotspot, refers to a large number of users' frequent stop-point in the trip process of urban residents, such as malls, supermarkets and other geographical locations.The urban hotspot area reflects the intensive area of urban residents' trip to a certain extent, which is of great guiding significance to the precise control of urban traffic.

The purpose of clustering algorithm analysis is to classify the data sets according to the characteristic attributes of the data objects, analyze the potential relation between the data, and obtain the semantic information of the data sets. Common clustering algorithms are divided into the following categories: clustering based on partitioning type, hierarchical clustering method, clustering method based on density type, clustering method based on grid type and model-based clustering method [7].

The basic idea of the hierarchical clustering algorithm[8] is to use the inter-cluster distance as the standard of the cluster similarity. When the distance between clusters satisfies the specific clustering rules, it is clustered into a new cluster.Its characteristics are: do not require the expected number of clusters expected to be divided, but through the hierarchical classification to the cluster.Hierarchical clustering can be divided into two categories according to their clustering order: AGglomerative NESting(AGNES) and DIvisive ANAlysis(DIANA).

This paper performs a clustering analysis of all user stop-points, uses AGNES clustering algorithm to analyze a large number of stop-point data, Therefore, it allows the parameter of space distance scale to be set, rather than the cluster numbers or the number of points gathering into a cluster,because we can not know and verify the accuracy of these parameters of the number of clusters or points of clusters in advance,so AGNES is more reasonable than other types of clustering algorithm.

For the same region, a large number of anchor samples appear on the map as the center-dense edge sparse characteristics, therefore, this paper uses the AGNES clustering algorithm with centroid distance to carry on the clustering analysis of a large number of users.The process of the AGNES algorithm based on centroid distance is:firstly, assume that there are n objects datasets, the n objects are treatedas n initial clusters,calculate the inter-cluster distance d of the n clusters and find the two clusters with the smallest distance, if the smallest distance between the two clusters meets the cohesive rule,then gather the two clusters into a new cluster.Then,calculate the inter-cluster distance d between the new cluster with other old clusters,update the distance between clusters,once again to find the smallest distance between two clusters among these clusters, if it meet the cohesion ruler, then merger the two,then calculate the inter-cluster distance between the new cluster with other old

clusters,update the distance d between clusters.And so on until the clustering stop condition is reached.The specific algorithm steps are as follows:

First, define each cluster as Cluster c, which is a collection of points,define the centroid of clusters as the point whose distance sum to all other points is smallest in the cluster,define the similar distance between two clusters as the spatial distance between the centroid points of the two clusters,define the radius of the cluster as the maximum distance among the centroid of the cluster and the other points in the cluster,define ClusterSet as a collection C of clusters,that is, the object of the algorithm calculation, but also the final output of the algorithm, as shown in equation (1).

$$C = \{c_i\}. \tag{1}$$

Step1.initialize each stop-point as a cluster c,and put it in the cluster collection C,so the size of collection C is equal to the number of stop-point when initialized.

Step2.traversing cluster collection ClusterSet,find the nearest cluster as Cluster A and Cluster B,pre-merge them to get Cluster AB,calculate the radius of the merged cluster AB,if it is less than the set threshold, merge them,and put Cluster AB as a new cluster into the CluserSet,remove the ClusterA and ClusterB to conduct the next step,otherwise, the algorithm ends and outputs the current cluster cluster as the final result.

Step3.repeat step 2 until you can not synthesize a new cluster.

Step4.After the above steps are completed, the clustering process is completed, the clustering results of the final urban hotspot area are in the ClusterSet object, each Cluster object represents the corresponding hotspot. The specific algorithm analysis steps are shown in Fig.1, the implementation of the algorithm for urban travel hotspots is shown in the pseudo code of Table 1.

Table 1. City Trip Hotspot Discovery Algorithm Pseudo Code

```
input:stop-point collection, S = {p_i | p_i · s_i = 1, p_i ∈ T_p}
output: hot spot clusters, C={ci}
initialization: ci←pi, for 1<=i<=S,C={ci}
Loop
(ci,cj)←FindClosestPair(C);
If    Radius(ci,cj)<=d
ci←Merge(ci,cj);
C←C-cj;
Else
Exit
End
```

Which FindClosetPair method is to find the nearest two clusters ci, cj in the cluster collection C;Radius (ci, cj) is to calculate the radius of the cluster mergedby ci and cj.Merge (ci, cj) is to merge clusters ci and cj.

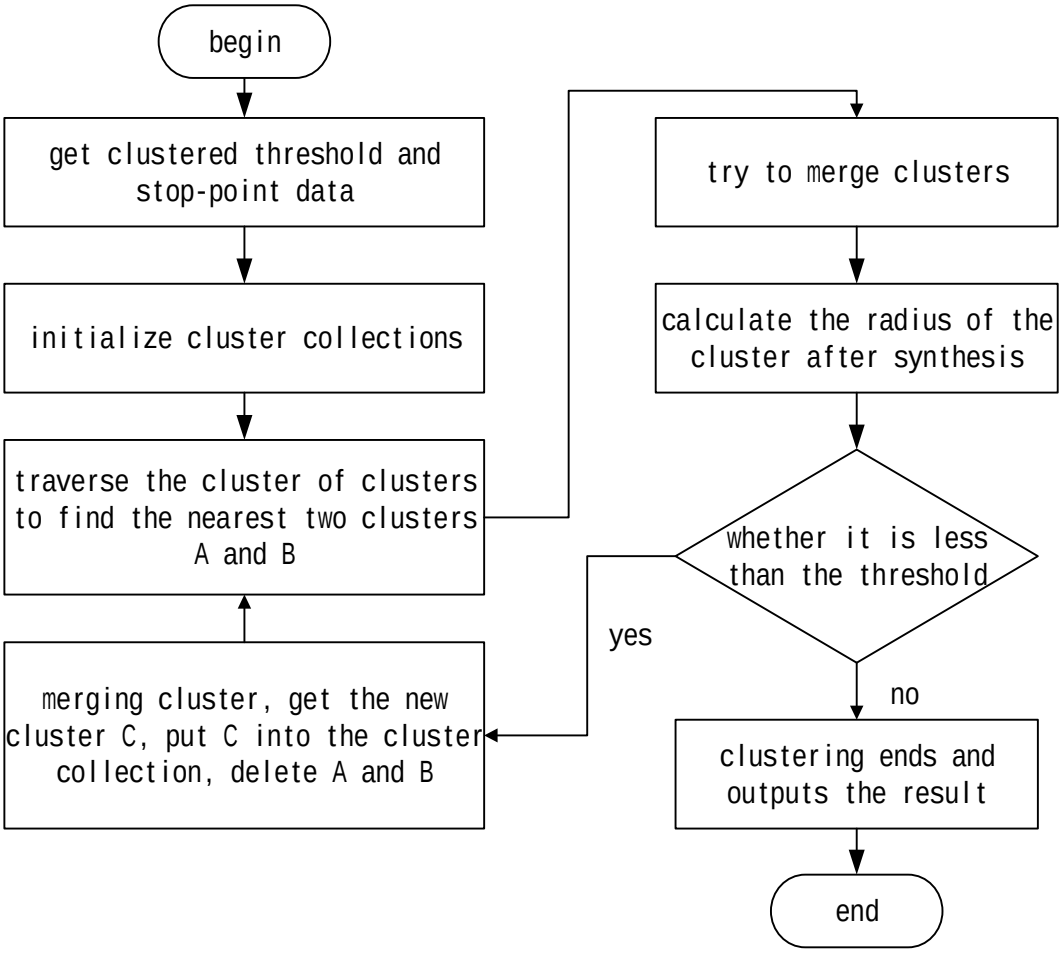

Fig.1. AGNES hotspot discovery process based on centroid distance

Case analysis

In this paper, through trip research based on mobile phone survey APP developed by research group, collected a total of more than 300 thousand track data from 129 volunteers in Wuhan China, part of the data shown in Figure 2.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	traceId	Latitude	Longitude	currentTime	speed	accuracy	adcode	accelerate x	accelerate y	accelerate z	accelerate xyz	altitude	bearing
2	305875	30.513279	114.344772	19:52:41	0	25	420111	0.002961814	0.003419876	-0.004239082	0.006199816	0	0
3	305876	30.513279	114.344772	19:52:43	0	25	420111	-0.018413723	0.011844158	-0.023480415	0.032104193	0	0
4	305877	30.513279	114.344772	19:52:48	0	25	420111	0.009797394	-0.003389835	-0.004715919	0.011389461	0	0
5	305878	30.513279	114.344772	19:52:53	0	25	420111	0.024651825	-0.013289928	0.010437965	0.029887887	0	0
6	305879	30.513279	114.344772	19:53:02	0	25	420111	0.013593614	-0.003455639	0.008728981	0.016520378	0	0
7	305880	30.513279	114.344772	19:53:03	0	25	420111	-0.000561118	0.014900684	-0.000205994	0.014912669	0	0
8	305881	30.513279	114.344772	19:53:09	0	25	420111	0.007661164	0.004004955	-0.007184029	0.011240256	0	0
9	305882	30.513279	114.344772	19:53:13	0	25	420111	0.011244118	0.011073112	0.018395424	0.024237072	0	0
10	305883	30.51320882	114.3447428	19:53:14	0.94	131		0.002309054	0.006482124	-0.004617691	0.008286901	22	188.1999969
11	305884	30.51322211	114.3447179	19:53:20	0.27	117		0.054771751	-0.013984203	-0.014677048	0.058403066	23	39.5
12	305885	30.51322266	114.3447176	19:53:26	0.602	198		-0.000274599	0.012175083	-0.012482643	0.017439164	24	41.29999924
13	305886	30.51328993	114.3446696	19:53:32	0.831	192		-0.001870573	-0.006516933	0.014318466	0.015842599	23	43.29999924
14	305887	30.51327501	114.3446796	19:53:37	1.36	186		0.036964715	-0.018132687	0.00939846	0.042231689	23	44.90000153
15	305888	30.51322428	114.3448315	19:53:49	0.827	229		-0.051434577	0.007339954	-0.047262192	0.070236068	23	48.90000153
16	305889	30.51324029	114.3447076	19:53:49	1.201	168		0.012056708	-0.039475918	-0.021780968	0.046670363	23	46.5
17	305890	30.51329508	114.3448503	19:53:56	2.301	194		-0.01728186	0.027519703	-0.082968712	0.089105575	23	50
18	305891	30.51333984	114.3447662	19:54:02	1.142	187		-0.021134675	-0.006900311	0.017457008	0.028267223	23	51.90000153
19	305892	30.51327365	114.3447512	19:54:07	1.126	215		0.006519556	0.013072968	0.00845623	0.016879423	23	52.90000153
20	305893	30.513279	114.344772	19:54:20	0	25	420111	0.029788375	-0.001455784	-0.005991936	0.030419892	0	0
21	305894	30.5132137	114.3447708	19:54:20	0.52	154		-0.092361569	-0.000864983	0.004824638	0.09249154	24	55.70000076
22	305895	30.5132468	114.3447862	19:54:33	0.645	145		0.01013422	0.006015778	-0.002943039	0.012147159	25	60
23	305896	30.51323134	114.3447561	19:54:39	0.36	240		0.004502356	0.006421089	0.016470909	0.0182426	26	61.40000153
24	305897	30.513279	114.344772	19:54:54	0	25	420111	-0.028235316	-0.021855354	0.014462471	0.038523404	0	0
25	305898	30.513279	114.344772	19:54:54	0	25	420111	-0.003978014	0.024552822	-0.020555496	0.032267539	0	0
26	305899	30.513279	114.344772	19:55:00	0	25	420111	-0.028023899	-0.004102707	0.00153923	0.02836442	0	0
27	305900	30.51324816	114.344719	19:55:05	1.04	147		0.008132994	0.019912243	0.01809597	0.028108845	28	352.1000061
28	305901	30.51326986	114.3447423	19:55:10	1.284	246		0.029021502	-0.008618832	-0.019774437	0.036160201	28	69
29	305902	30.513279	114.344772	19:55:21	0	25	420111	0.028814256	-0.006354332	0.015677452	0.033412893	0	0
30	305903	30.513279	114.344772	19:55:26	0	25	420111	0.029759467	-0.024071693	0.022320747	0.044309007	0	0
31	305904	30.513279	114.344772	19:55:31	0	25	420111	-0.038610041	0.031173706	-0.021551132	0.054101631	0	0
32	305905	30.51329861	114.3445635	19:55:35	1.49	148		0.003600597	-0.005080223	-0.006864548	0.009267954	31	92.69999695

Fig.2.Part of the travel survey data set

In this paper, the AGNES trip hotspot discovery algorithm based on centroid distance is used to perform a clustering operation for a total of 203 stop-points generated by the user during the system travel investigation, the clustering threshold of the algorithm is 500 m, and generates a total of 46 Hot spots; where the heat value of the top three hot spots as shown in Table 2.

Table 2 hot spot value of the hot area

No.	latitude	longitude	Heat value
1	30.510942	114.343353	56
2	30.514786	114.344236	42
3	30.601978	114.358027	27

The heat value is the total number of people staying in this area. It shows that the hot spots areas are basically in line with the actual results of this travel survey.

Conclusions

In this paper, an AGNES algorithm based on centroid distance is proposed to extract the hot spots in urban residents combining the stop-point of the mobile travel survey, the method takes reasonable consideration of the unknown degree of the density of the hot spot area, and obtains the travel heat value of the urban residents according to the residence intensity of the travel area. The validity of the algorithm is verified by the data based on the mobile travel survey.

References

[1] Mao Haixiao.*A Study on the Traveling Characteristics of Chinese Urban Residents*[D].Beijing:Beijing university of technology,2005.

[2] Hariharan R, Toyama K. Project Lachesis: *Parsing and Modeling Location Histories*[M]// Geographic Information Science. Springer Berlin Heidelberg, 2004: 106-124.

[3] Calabrese F, Mi D, Lorenzo G D, et al. Understanding Individual Mobility Patterns from Urban Sensing Data: A Mobile Phone Trace Example[J]. Transportation Research Part C Emerging Technologies, 2013, 26(1): 301–313.

[4] Angelakis V, Gundlegård D, Rydergren C, et al. Mobility Modeling for Transport Efficiency: Analysis of Travel Characteristics Based on Mobile Phone Data[C]//Netmob 2013-Third International Conference on the Analysis of Mobile Phone Datasets, May 1-3, 2013, MIT, Cambridge, MA, USA. 2013.

[5] Hu Zhongwei,Deng Xiaoyong,Guo Jifu,etc..Analysis on Resident Trip Demand Characteristics Based on Mobile Phone Location Data[J].The 8th China Intelligent Transportation Annual Conference Proceedings--Rail Transit, 2013.

[6] Kim Y, Pereira F C, Zhao F, et al. Activity Recognition for a Smartphone Based Travel Survey Based on Cross-User History Data[C]// International Conference on Pattern Recognition. IEEE, 2014: 432-437.

[7] Guo Junhua.***Research on Clustering Analysis in Data Mining***[D]. Wuhan:Wuhan University of Technology, 2003.

[8] Patel S, Sihmar S, Jatain A. A Study of Hierarchical Clustering Algorithms[C]// International Conference on Computing for Sustainable Global Development. IEEE, 2015:537-541.

[9] Murtagh F, Contreras P. Algorithms for Hierarchical Clustering: An Overview[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012, 2(1): 86-97.