2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovations

Clustering Large-Scale Origin-Destination Pairs: A Case Study for Public Transit in Beijing

Miao Li, Beihong Jin, Hongyin Tang and Fusang Zhang
State Key Laboratory of Computer Sciences, Institute of Software, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China

***Abstract*—With the extensive collection of various trajectories, a lot of trajectory mining methods have been developed and brought into effect in different applications. The same is true for trajectory clustering. It enables the construction of diverse applications (e.g., mobile social networks) and can promote the intelligence of existing services (e.g., optimizing public transit). In the paper, we propose a three-phase clustering strategy ODTC (Origin Destination pair oriented Trajectories Clustering) for the massive trajectories in the form of OD (Origin Destination) pairs and demonstrate the impact of trajectory clustering on evaluating and adjusting public transit operations. In our ODTC strategy, trajectories are partitioned in the first phase by coarse-grained clustering, reflecting an idea of divide and conquer. While during the second phase of fine-grained clustering, we model the relations of OD pairs as a sparse graph where the spatial and temporal features as well as the constraints of road networks are integrated into the similarity of trajectories. Then we apply a spectral clustering algorithm on the graph to capture clusters. In particular, in the third phase, we borrow the idea from text data mining and give a feasible method to mine the semantics of clusters. As a case study, we perform ODTC on the large-scale trajectories from the Beijing Public Transport Group. From the clustering results, we can observe the mobility patterns of bus passengers. Further, we exploit the clustering results to discover the dynamics of bus operations, evaluate the bus lines and provide support for making the decisions on bus operations.**

***Keywords*—Trajectory, Origin-Destination Pair, Clustering, Intelligent Transportation System, Data Mining**

I. Introduction

Trajectory data of moving objects (such as people, vehicles, animals and natural phenomena) are a series of location records over time [1]. The most common ones are chronological location data (e.g., GPS data) recorded by positioning receivers which move along with the moving objects. The other trajectory data include signaling sequences generated by mobile phones carried by the moving objects, and OD (Origin Destination) pairs generated while the people swipe the smart cards on buses or subways. Trajectory data imply the spatial and temporal features of object movements, and various data analytics technologies have been applied to observe mobility of moving objects [2][3].

Clustering is to group data in clusters, each of which is as distinct as possible from the others. Understanding and utilizing the mobility reflecting in the clustering results enable many interesting applications to be constructed. For example, air mass back trajectories can be clustered, using the latitude and longitude of a trajectory along with the chemical species (i.e., carbon monoxide) in the air mass as clustering variables [4]. From the clustering results, meteorologists can know the source region of certain chemical species that the air mass contains, and further recognize seasonal difference of the source region of the air mass. By clustering the common sub-trajectories of animals (e.g., elks) [5], zoologists can discern whether the herd has the habit of moving together and identify the impacts of the road traffic flows on the movements and distribution of animals. On the other hand, we note that clustering the trajectories of public transport passengers in a city can understand the passenger mobility. More specifically, the bus passengers can be clustered by the stops they get on and off the buses, as a result, the passengers can be classified into commuters, regular OD passengers, habitual time passengers, and irregular passengers [6]. As thus, the preferential policies for classified passengers can be designed. Moreover, by clustering the bus passengers, the strangers who can meet each other at the stops or on the buses, that is, "familiar strangers" can be found [7]. Exploiting the identified familiar strangers, social tools can be built to establish the connection between strangers and extend the scope of making friends. As for vehicle trajectories, the clustering can be used to predict the vehicle collision [8]. If using trajectories of sharing-bikes, planning the bike lanes is also feasible [9].

So far, existing data clustering approaches, including k-means [4], spectral clustering [8], hierarchical clustering [10], density-based clustering [5], model-based clustering [11] [12], have been applied to trajectory data. While employing the existing approaches, several key points should be taken into account, including (1) how to extract the features of different trajectories (e.g. trajectories with widely varying lengths, trajectories on road networks) [13], (2) how to fuse the spatial and temporal factors in clustering [14], (3) how to exploit the semantics of trajectories [7] and (4) how to show the practical values of clustering results [15].

However, facing large-scale trajectories, the challenges mainly come from the following two aspects. (1) The pre-processing time of trajectories is non-negligible. For example, in general, the clustering procedure needs the similarity between objects being clustered, which lays the foundation for performing different clustering algorithms. If the number of trajectories is huge, it will take a lot of time to calculate the similarities between any two trajectories, which greatly decrease the efficiency of clustering. (2) There is a gap between clustering quality and application requirements. The quality of clustering results can be measured by multiple metrics, e.g., the number of clustering, inter-cluster cohesion degree,

978-1-5386-9380-3/18/$31.00 ©2018 IEEE
DOI 10.1109/SmartWorld.2018.00137

intra-cluster dispersion degree and semantics of clustering. Too many or too few clusters, low inter-cluster cohesion, high intra-cluster dispersion and lack of clustering semantics will bring difficulties to applications of clustering results.

Facing the above challenges, in this paper, we propose a three-phase clustering strategy named ODTC (Origin Destination pair oriented Trajectories Clustering) for large-scale OD pairs, where three phases refer to coarse-grained clustering, fine-grained clustering and semantics mining.

While compared to the existing work, our work has the characteristics as follows. First, for dealing with large-scale trajectory processing, we perform coarse-grained clustering on all the trajectories, and then perform fine-grained clustering on the results of coarse-grained clustering. Such a strategy of the "divide and conquer" allows users to carry on the second clustering on some selected results of coarse-grained clustering. Thus, users can get the clustering results without waiting for whole trajectory processing to end and quickly find what their applications need from the clustering results. Second, for the trajectories in the form of OD pairs, we model OD pairs and their relations as vertices and edges in a graph, in particular, the spatial and temporal features as well as the constraints of road networks are integrated into the similarity of trajectories, i.e., the weights of edges. The ODTC strategy is to perform the clustering on the above graph. Meanwhile, we borrow the idea from text mining and give a feasible method to mine the semantics of clusters. Third, to bridge the gap between application requirements and a clustering strategy, we demonstrate the impact of trajectory clustering on public transit operations. We collect the smart card data of passengers from the Beijing Public Transport Group and extract from them the OD pairs of bus passengers. After applying the ODTC strategy to OD pairs of bus passengers, we analyze the mobility of bus passengers from clustering results. Further, we evaluate the transport capacity of bus lines and make some suggestions on improving bus transit operations. At last, we note that the OD pair is an essential constituent of a trajectory and current sensing technology makes OD pairs relatively easy to be sensed and collected, therefore the ODTC strategy with the OD pairs as input can be applicable to various scenarios. For example, if the OD pairs of passengers who take the other public transportation (i.e., subway and taxi) are available, the solution in the paper can get a big picture of entire public transit.

The rest of the paper is organized as follows. Section II gives the description of the ODTC strategy. Section III outlines the experimental evaluation of the ODTC strategy. Both Section IV and Section V analyze the clustering results of OD pairs of bus passengers in Beijing, the former involves in examining the passenger mobility and the latter concerns optimizing public transit operations. Finally, Section VI concludes the paper.

II. Clustering Origin-Destination Pairs

In general, a trajectory of a passenger at least consists of a passenger's ID, and origin and destination information, where both the origin and destination information can be represented as a triple, i.e., (lat, lon, t), indicating latitude, longitude and a timestamp, respectively. An OD pair is a simplified trajectory of a passenger, only containing the boarding and disembarking information. It is denoted as$(O, D) = (O.lat, O.lon, O.t, D.lat, D.lon, D.t)$. From the definition, the OD pairs can be extracted from the passenger trajectories directly.

For such OD pairs, we propose a three-phase clustering (i.e., coarse-grained clustering, fine-grained clustering, and semantics mining) approach to reveal the mobility patterns of passengers.

A. Identifying Coarse-grained Clusters

Given an OD pair, we define the length of the OD pair as the Euclidean distance between the origin and destination. We note that the length difference between OD pairs is an important indication of diversified passenger demands of taking buses, further reflects the diversity of mobility patterns. Therefore, we design the following coarse-grained clustering.

We first sort all the OD pairs in ascending order of the length of the OD pair. Then we group the ordered OD pairs into several sets where the length difference of any two OD pairs in a same set is less than the predefined threshold w. Finally, we sort all the sets according to the descending order of the number of elements in the set, and mark these sets as resulting coarse-grained clusters. Algorithm 1 gives the pseudo code of coarse-grained clustering.

Algorithm 1: Coarse-grained Clustering

```
Input: PS: OD pairs, w;
Output: CCS: Coarse-grained clusters;
Sort PS;
ccs_index ← 0;
while count(PS) > 0 do
    pre ← 0, post ← 0, target ← 0;
    for i = 1; i ≤ count(PS); i++ do
        if length(PS[post]) − length(PS[pre]) > w then
            target ← pre;
        else
            pre ← pre + 1;
        end
        post ← post + 1;
    end
    CCS[ccs_index] = {PS[target], ..., PS[target+post−pre]};
    ccs_index ← ccs_index + 1;
    Remove all OD ∈ CCS[ccs_index] from PS
end
return CCS;
```

As a result, each coarse-grained cluster contains a certain size of OD pairs which have similar lengths or whose length differences are within w. By default, w is set to 2 km.

B. Identifying Fine-grained Clusters

From the characteristics of OD pairs, we can derive that the OD pairs have many-to-many relations such as distance, similarity, etc. Therefore, based on the OD pairs from the same coarse-grained cluster, we construct a graph for OD pairs (called the similarity graph) to capture the closeness between OD pairs. Then we apply the spectral clustering algorithm to find the clusters of OD pairs.

The similarity graph is in essence an ε-neighborhood weighted graph, where each node denotes an OD pair. For two OD pairs, denoted by x and y, there is an edge between them if the similarity is larger than ε. Moreover, the similarity greater than ε is used as the weight of the edge. We define the similarity (denoted as ODS) as the weighted sum of temporal similarity and spatial similarity in the following form:

$$ODS(x, y) = \alpha \times TS(x, y) + \beta \times SS(x, y) \quad (1)$$

where $TS(x, y)$ and $SS(x, y)$ denote temporal and spatial similarity taking α and β as their weights, respectively.

What we highly desire is that the boarding times of the OD pairs in a cluster are as close as possible to each other and the temporal similarity between two OD pairs decreases with the increase of difference of the boarding times. Therefore, we apply a Gaussian (RBF) kernel to transform the time difference and define the temporal similarity as

$$TS(x, y) = \exp(-\frac{\eta^2}{2 \times \rho^2}) = \exp(-\frac{|x.O.t - y.O.t|^2}{2 \times \rho^2}) \quad (2)$$

where ρ is a free parameter representing the width of the Gaussian kernel. By default, we set ρ to 18. Thus, TS is greater than 0.6 when the time difference between two origins is less than 18 min and approaches 0 when the time difference is greater than 60 min. Such setting conforms to our life experience. Fig. 1 shows curve of $TS(x, y)$ function while assigning ρ to 18.

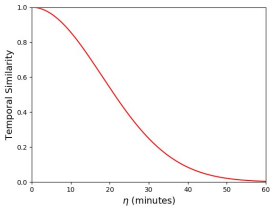


Fig. 1. Temporal similarity.

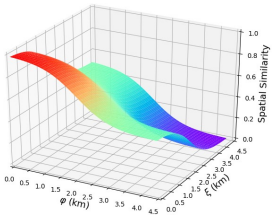


Fig. 2. Spatial similarity.

On the other hand, if for two OD pairs, their origins are geographically close to each other and so are their destinations, then these two OD pairs are considered to be similar in spatial dimension. We use a Logit function to measure such spatial similarity. $SS(x, y)$ is formulated as

$$\begin{aligned} SS(x, y) &= \lambda \times \frac{e^{-\pi_o(\phi - \tau_o)}}{1 + e^{-\pi_o(\phi - \tau_o)}} + \mu \times \frac{e^{-\pi_d(\xi - \tau_d)}}{1 + e^{-\pi_d(\xi - \tau_d)}} \\ \phi &= dist(x.O.lat, x.O.lon, y.O.lat, y.O.lon) \\ \xi &= dist(x.D.lat, x.D.lon, y.D.lat, y.D.lon) \end{aligned} \quad (3)$$

where $dist(lat1, lon1, lat2, lon2)$ denotes the shortest distance between two geographic locations represented by $(lat1, lon1)$ and $(lat2, lon2)$ through applying a map matching algorithm and the Dijkstra algorithm on the road network, and ϕ and ξ denote the distance between origins and distance between destinations on a road network, respectively.

In addition, λ and μ in Eq.3 are the parameters representing the weights of origins and destinations, respectively. If setting λ or μ to 0, we can obtain the radial clusters whose origins or destinations are dispersive.

After we set λ and μ to 0.5, π_o and π_d to 2.0, and τ_o and τ_d to 2.0, the curve of $SS(x, y)$ function is shown in Fig. 2. Fig. 2 shows that when ϕ and ξ are greater than 4.5, the spatial similarity tends to be 0.

From the above definition of the similarity between two OD pairs, we can get that the larger ODS value means the closer spatial and temporal relation between the OD pairs. Moreover, ODS is symmetric and non-negative, i.e., given two OD pairs x and y, we have $ODS(x, y) = ODS(y, x)$ and $ODS(x, y) > 0$.

Given n OD pairs, to divide the OD pairs into several clusters such that OD pairs in the same cluster are similar and OD pairs in different clusters are dissimilar to each other, we build the similarity matrix $\boldsymbol{S_{n \times n}}$ for the similarity graph where the value of element s_{ij} is $ODS(i, j)$, i.e., the similarity of OD pair i and j. And then, we apply a spectral clustering algorithm to $\boldsymbol{S_{n \times n}}$. In detail, we first build the Laplacian matrix $\boldsymbol{L = D - S}$ where $\boldsymbol{D}$ is a diagonal matrix and $d_{ii} = \sum_j s_{ij}$. Next we calculate the eigenvectors corresponding to the top-g smallest eigenvalues of the symmetric normalized Laplacian defined as $\boldsymbol{D^{-1/2} L D^{-1/2}}$ and then form the matrix $\boldsymbol{F}$ with the above eigenvectors as columns and further normalize the rows. Finally, we regard each row in $\boldsymbol{F}$ as a sample and cluster these n samples with the k-means algorithm into k clusters. Here, the values of g and k have to be designated in advance.

C. Inferring Semantics of Clusters

After we find the clusters of OD pairs, we also hope to mine the semantics of these clusters. Only in this way, we can understand the movements of bus passengers and further the mobility tendency of city residents and utilize for applications.

For this purpose, we borrow the idea of the document topic identification and mine the semantics of clusters.

For each OD pair in a cluster, we collect the information about POIs (Points of Interest) close to the origin and the destination to be two profiles, employ a word segmentation tool Jieba[1] to divide POI information into terms, and aggregate the profiles of origins and destinations in a cluster to be two different POI documents. Thus, we get a set of POI documents.

For this set of POI documents, we calculate the TF-IDF value of each POI term, build the matrix $\boldsymbol{V_{m \times k}}$ where a column denotes a POI document and a row denotes the terms in it, and v_{ij} is TF-IDF value of the i-th term in the j-th document. Then, we apply a non-negative matrix factorization method i.e., CNMF (Constrained Non-negative Matrix Factorization) to factorize the matrix $\boldsymbol{V}$ into two matrices $\boldsymbol{W}$ and $\boldsymbol{H}$, where $\boldsymbol{H}$ is called topic-document matrix and a column in $\boldsymbol{H}$ gives the topic distribution in a document, and $\boldsymbol{W}$ is called term-topic matrix and a column in $\boldsymbol{W}$ gives the terms attached to a specific topic. More specifically, $\boldsymbol{V_{m \times k} = W_{m \times h} H_{h \times k}}$, where m is the number of terms in the set of documents and h is the number of topics and should be less than k. Finally, for a cluster, let column r be the corresponding document column in $\boldsymbol{H}$ and h_{qr} has the maximum value in column r, we can get terms corresponding to top-l largest values from column q

[1]https://github.com/fxsjy/jieba

of matrix $\boldsymbol{W}$ and take these terms as the semantic description of the cluster. By default, l is set to 15.

In summary, we present an OD pair oriented three-phase clustering approach which combines the characteristics of trajectory data and application goals. During the coarse-grained clustering, we differentiate the various mobility patterns at different granularity of distances. During the fine-grained clustering, we define the similarity of OD pairs and cluster OD pairs within each coarse-grained cluster. In our approach, each cluster contains some close related OD pairs, corresponding to the passenger mobility trends in cities. Finally, we identify the semantics of a cluster by analyzing POIs at the surrounding of the origins and destinations of OD pairs in a cluster.

III. Experimental Evaluation

In this section, we conduct experiments on real-world smart card data from the Beijing Public Transport Group, and by experimental results we illustrate the rationalization of method selection in the ODTC strategy.

A. Selecting the Fine-grained Clustering Method

The alternative methods for fine-grained clustering are the spectral clustering, the Louvain algorithm, DBSCAN, and the hierarchical agglomerative clustering (HAC) algorithm. Table I lists the parameter setting in these clustering methods.

TABLE I. Parameters setting in clustering methods

Method	Parameters
Spectral clustering	$g = k = 155$
Louvain	None
DBSCAN	$eps = 0.03, n\_neighbors = 5$
HAC	$n\_clusters = 155$

We collect the passenger trajectories on Oct. 31, 2016 and obtain 5,626,571 OD pairs. Then, we conduct coarse-grained clustering on these OD pairs and obtain 52 coarse-grained clusters. For experiment convenience, we random choose 10 clusters from all the coarse-grained clusters and conduct fine-grained clustering for each coarse-grained cluster.

We adopt the following metrics to measure the quality of fine-grained clustering results.

- Number of resulting clusters (NRC).
- Mean time difference (MTD): MTD is the mean of Xs of all the clusters, where X is the mean time difference of the origins of any two OD pairs in a resulting cluster.
- Mean distance (MD): Let Y_1 denote distance difference of the origins in an OD pair and Y_2 denote distance difference of the destinations. Let Y be the mean of Y_1 and Y_2 of any two OD pairs in a resulting cluster. MD is the mean of Ys of all the clusters.
- Mean length difference (MLD): MLD is the mean of Zs of all the clusters, where Z is the mean of length difference of any two OD pairs in a resulting cluster.
- Calinski-Harabaz Index (CHI [16]): CHI is ratio of the inter-clusters dispersion mean and intra-cluster dispersion. The larger CHI is, the better the quality of resulting clusters is. Let n be the total number of OD pairs and k be the number of resulting clusters. CHI is defined in Eq.4. $Tr(.)$ denotes the trace of matrix, $\boldsymbol{B}$ is the covariance matrix of inter-clusters data, and $\boldsymbol{W}$ is the covariance matrix of intra-cluster data.

$$\text{CHI} = \frac{Tr(\boldsymbol{B})}{Tr(\boldsymbol{W})} \times \frac{n-k}{k-1} \tag{4}$$

Table II lists the metric values obtained from the resulting clusters, which guides us to the choice of the spectral method as the fine-grained clustering method.

TABLE II. Quality of clusters of fine-grained clustering methods

Method \ Metrics	NRC	MTD	MD	MLD	CHI
Spectral clustering	155	48.33	1.40	0.49	363.26
Louvain	511	58.71	3.59	0.49	121.31
DBSCAN	314	54.59	2.33	0.19	4.44
HAC	155	53.89	2.82	0.54	6.352

B. Selecting the Semantics Mining Method

Two competitive methods for semantics mining are CNMF and LDA (Latent Dirichlet Allocation).

We execute coarse-grained and fine-grained clustering over the OD pairs on Oct. 31, 2016 and obtain 8060 fine-grained clusters. For these 8060 clusters, we form 16,120 POI documents, each cluster with two documents (one for origins and another for destinations in a cluster). We set the number of topics h to 20 according to the categories of POIs. In the term-topic matrix of CNMF and LDA, we choose top-15 largest terms for each topic.

We adopt the following two metrics to measure the quality of topics generated by different mining methods.

- topic coherence [17]
  coherence$(t) = \sum_{j=2}^{l}\sum_{i=1}^{j-1} \log \frac{C(v_i,v_j)+1}{C(v_j)}$, where t denotes a topic, v_i and v_j denote top i-th and top j-th terms in l most probable terms describing the topic t, $C(v_i, v_j)$ counts the number of documents containing term v_i and v_j, and $C(v_i)$ counts the number of documents containing term v_i.
- topic independence [18]
  independence$(t_i, t_j) = 1/l \times \sum_{z=1}^{l} E_z(t_i, t_j)$, where $E_z(t_i, t_j) = |T_{i,d} \cap T_{j,d}|/|T_{i,d} \cap T_{j,d}|$, t_i and t_j denote topics, l denotes the number of terms describing a topic, $T_{i,d}$ denotes a set of top-d terms describing topic t_i.

The above metrics reflect our requirements of topics, that is, terms in a same topic represent similar semantics and are expected to occur in a same document as much as possible, however terms under different topics represent different semantics and should be as different as possible.

For 20 topics generated by CNMF and LDA, respectively, we calculate the means of their coherence and independence. The results are listed in Table III.

TABLE III. Semantic quality of topics

Method \ Metrics	Independence	Coherence
CNMF	0.4485	-65.4755
LDA	0.0157	-45.0636

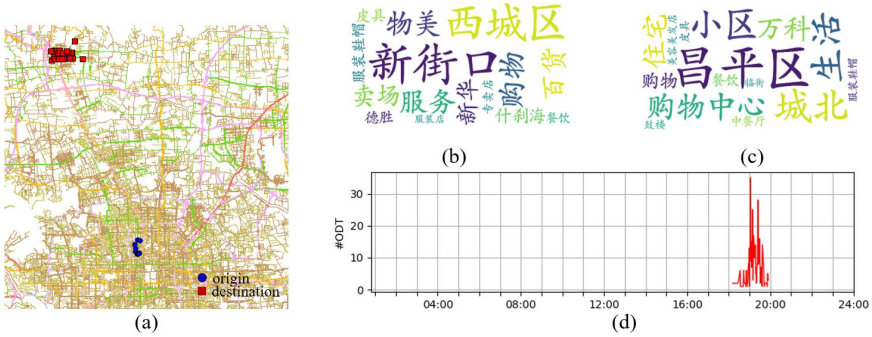


Fig. 3. Xinjiekou→Changping District (19:00-20:00)

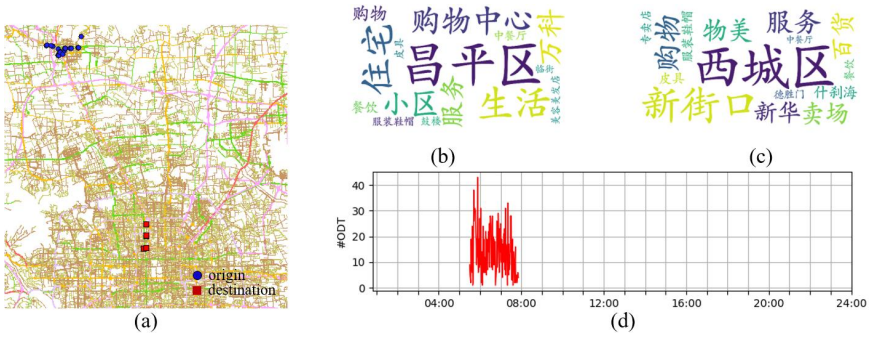


Fig. 4. Changping District→Xicheng District (06:00-07:30)

The experimental results show that topics from CNMF are superior to ones from LDA. By observing the generation processes of two methods, we find that LDA directly employs the frequency of terms as a feature and the top few terms of topic are often the ones which occurs in all documents, e.g., Beijing, company, etc. These terms decrease the distinction degree of different topics and pose negative influence on topic coherence. However, in the CNMF method which adopts TF-IDF as a feature, the weights of these terms will decrease, and semantic distinctive terms are more likely to rank high.

IV. Analyzing Passenger Mobility

To observe the passenger mobility, we conduct multiple experiments on the 36,901,893 passenger trajectories from Oct. 31, 2016 to Nov. 6, 2016 while setting different parameters for ODS. From the clustering results, we find different passenger mobility patterns, including wired patterns and radial patterns. Here, the wired pattern refers to such a cluster that origins are close together and so are destinations. And the radial pattern is a cluster that origins are close together and destinations are scattered.

A. Wired patterns

In the first experiment, we set α=1, β=1, λ=1, μ=1, and obtain 8525 clusters. From the resulting clusters, we choose 4 typical wired patterns and show them in Figs. 3-6. Fig. 3-6(a) shows the geographical locations of origins and destinations in a wired pattern, where we label the origins and destinations with the blue circles and red squares, respectively. Fig. 3-6(b) and (c) are the word clouds where the words come from terms representing the semantics of origins and destinations, respectively, giving users an indication of what the origins and destinations are about and helping users to comprehend the wired patterns. Fig. 3-6(d) shows the temporal distribution of passengers between these two areas.

Fig. 3(a) shows a wired pattern from an urban center to a northwestern area. From the word clouds in Figs. 3(b)-3(c) and temporal information in Fig. 3(d), we find that passengers who usually get off work or finish shopping, travelling from Xicheng District to Changping District, give rise to this pattern. Fig. 4 shows a wired pattern of the similar OD pairs but in opposite direction. The semantics expressed via the word cloud clearly shows the function of this northwestern area, i.e., Changping District is mainly a residential area in Beijing. The pattern in Fig. 5 shows that passengers move from Beijing Railway Station to Tongzhou District from 18:00 to 19:00. According to the word clouds, we infer that these passengers go back home at Tongzhou District, which coincide with the nature of most residential areas. Using our ODTC strategy, we can discover passenger mobility patterns not only in the morning or evening peaks but also during other time periods. Fig. 6 shows a wired pattern from 13:00 to 14:00. It indicates that a large group of passengers move from Chaoyang District to Beijing West Railway Station at noon.

In the second experiment, we set α=0, β=1, λ=1, μ=1. In other words, the similarity of two OD pairs is equal to their spatial similarity in this setting. The resulting clusters show the

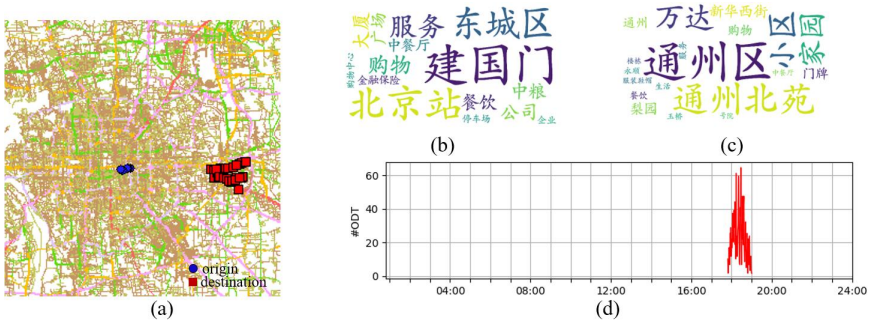


Fig. 5. Jianguomen→Tongzhou District (18:00-19:00)

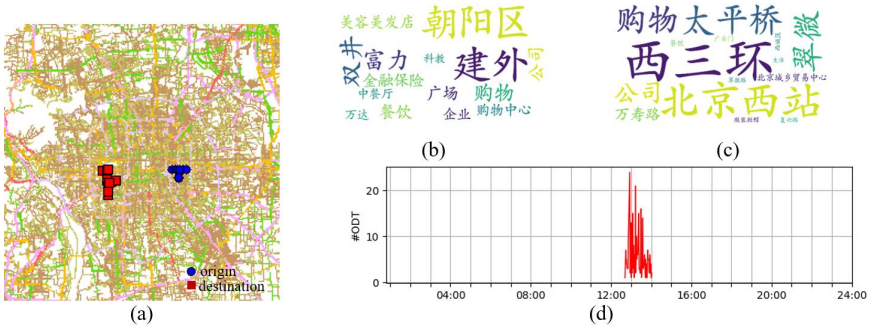


Fig. 6. Chaoyang District→Cuiwei (13:00-14:00)

overall trend of mobility of passengers in Beijing. We name these patterns after the semantics of origins and destinations. We show 15 typical resulting clusters in Fig. 7 and list some of their properties in Table IV.

From Table IV, we can find that the pattern between Tongzhou and Beijing Railway Station has more trajectories than others. Moreover, several railway stations in Beijing play an important role in bus transportation as half of the patterns are related to these railway stations.

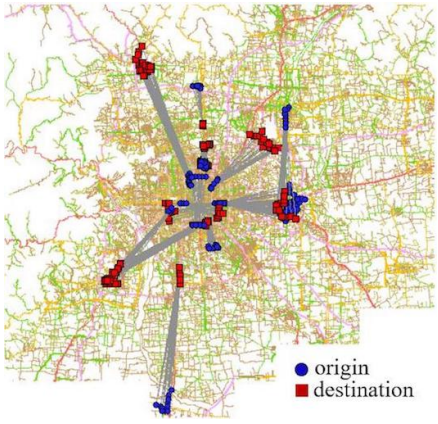


Fig. 7. Visualization of wired patterns

Also from the second experiment, we can obtain the clusters of short-distance trajectories (0.4-2.4 km). In Fig. 8, different colors are used to represent the resulting clusters in descending order of the number of short-distance trajectories in a cluster. For example, the black area denotes the area with the maximum number of short-distance trajectories and the light blue area (e.g. the left-most cluster) denotes the area with the 12th highest area. The black area indicates that passengers around Beijing West Railway Station take the most of short travelling than passengers in other areas.

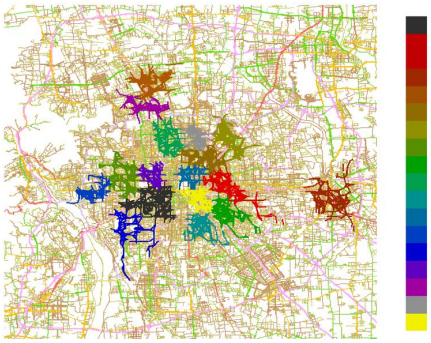


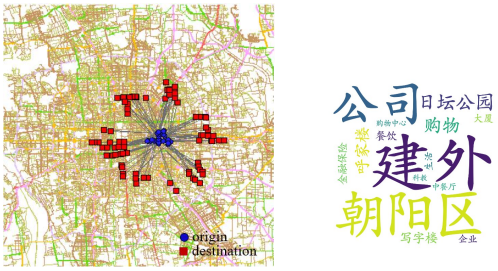
Fig. 8. Distribution of short-distance trajectories in Beijing

B. Radial patterns

In the third experiment, we set α=1, β=1, λ=1, μ=0. That is, we only focus on the similarity of origins when calculating the spatial similarity between two OD pairs. From the resulting clusters, we find the radial patterns, i.e., a large number of passengers moving from a particular area (called the

TABLE IV. Some selected wired patterns

Wired Pattern No.	Wired Pattern	Num of Trajectories	Num of Bus Lines
1	Tongzhou District Government → Beijing Railway Station	56270	10
2	Beijing Railway Station → Tongzhou District	53473	12
3	Beijing North Railway Station → Changping Railway Station	57202	7
4	Daxing District → Gu'an County	18141	2
5	Shunyi District → Tongzhou District	14582	1
6	Beijing West Railway Station → Fangshan District	59223	6
7	Dongzhimen Subdistrict → Beijing Capital Internatonal Airport	27796	4
8	Beijing South Railway Station → Fangshan District	13245	2
9	North gate of Olympic Forest Park → National Stadium	9131	14
10	National Stadium → North gate of Olympic Forest Park	6197	13
11	Fangzhuang Area → Jiugong Area	9985	3
12	Jiugong Area → Fangzhuang Area	7898	3
13	Beijing West Railway Station → Jinsong Subdistrict	3746	2
14	Chinese Aviation Museum → Tiantongyuan North Station	10481	3
15	Beijing Railway Station → Beijing West Railway Station	6134	7



(a) A radial pattern (b) Word cloud

Fig. 9. Visualization of a radial pattern.

(a) Cores in the morning peak (b) Cores in the evening peak

Fig. 10. Cores of radial patterns.

core hereinafter) to the surrounding areas. Fig. 9(a) gives the visualization of a cluster in the evening peak (16:00-20:00), where all the blue circles, each circle denoting an origin, and the other red squares denote the destinations. Company, Mall, Ri Tan Park and etc. are key words in the semantics of core in Fig. 9(b). From this radial pattern, we can know that many people need to take buses from the core to the surrounding areas. And we can infer that these passengers may work, go shopping, or take a tour at this area, and then go back to their residences. Besides, we are also able to know the interested destinations of the passengers.

Further, the cores of radial patterns are shown in two figures: Fig. 10(a) for the morning peak and Fig. 10(b) for the evening peak. Fig. 10 shows that Beijing Railway Station (shown in dark green triangles) and the Panjiayuan area (shown in dark red triangles) are the areas where the volume of departing passengers keeps high, no matter whether in the morning or the evening peak. Such observations exactly match the reality in Beijing, because at Beijing Railway Station there are always a lot of departing passengers, and a large antique market and a glasses market located in Panjiayuan area bring the heavy crowds. In addition, we observe that the Daxing district as a core of the radial pattern appears in the morning peak (shown in black triangles) and disappears in the evening peak. The reason behind is that the Daxing district in the southwest of Beijing is a residential area and far away from downtown. A large number of people leave the area for work only in the morning.

V. Optimizing Public Transit Operations

A. Planning inter-zonal buses

From clustering results (when α=1, β=1, λ=1, μ=1), we can analyze the time-variation characteristics in wired patterns.

Taking the clusters shown in Figs. 3-4 as examples, we observe that one wired pattern from Xicheng District to Changping District occurs in the evening peak (starting around 18:30 and lasting about 1.5 hours) and another one from Changping District to Xicheng District occurs in the morning peak (starting around 5:30 and lasting about 2.5 hours). While taking the time-variation characteristics into consideration, we suggest adding inter-zonal buses from Xinjiekou (the largest word in the word cloud, also a place name in Xicheng District) to Changping District at around 19:10 and ones in opposite direction at around 6:00. Similarly, from the wired pattern shown in Fig. 6, we suggest adding inter-zonal buses from Jianwai to Beijing West Railway Station at 13:00.

B. Adjusting bus lines

For radial patterns (when α=1, β=1, λ=1, μ=0), we note that the coverage of bus lines is very different. By analyzing the number of bus lines that a radial pattern covers, we can find areas with insufficient or redundant transport capacity, and further suggest the Beijing Public Transport Group adjust the bus lines and make the public transit operations with high efficiency. Table V shows some properties of selected radial patterns. From the table, it is obvious that the bus lines that pattern No. 5 covers are much more than those of pattern Nos. 2-4, but the number of trajectories is much less than ones of those patterns.

TABLE V. Some selected radial patterns

Radial Pattern No.	Peak Hours	Original Area	Num of Trajectories	Num of Bus Lines
1	M	Dongcheng Distric	192608	216
2	M	Beijing Railway Station	161116	191
3	M	Beijingxi Railway Station	146077	183
4	E	Beijing Railway Station	123392	163
5	M	Fangshan District	118647	211
6	E	Beitaipingzhuang	113189	160
7	M	Beijingnan Railway Station	113103	190
8	M	Xicheng District	112950	186
9	M	Zhongguancun	139597	155
10	E	Olympic Park	108886	198

TABLE VI. Trajectories from Tongzhou to Beijing Railway Station of different bus lines

Bus line No.	668	626	667	666	647
Num of Trajectories	5002	8920	11403	54	8511
Bus line No.	809	808	807	312	648
Num of Trajectories	3971	5159	13248	70	25

Similarly, pattern No. 9 is covered by 155 bus lines which has less number of bus lines than pattern No. 7 and No. 10, but it has a higher volume of trajectories. The paradox phenomenon is supposed to be attributed to the fact that the managers do not fully understand the passenger mobility patterns, leading to an excess or shortage of regional bus transport capacity. Therefore, we suggest reducing the number of bus lines that pattern No. 5 covers and adding the number of bus lines for pattern No. 9, so that the operation capacity of the bus system can be improved.

In addition, the clustering results (when α=0, β=1, λ=1, μ=1) can be a basis for optimizing the public transit system. By analyzing the number of trajectories in each cluster and the bus lines covering that cluster, we can find the inappropriateness of public transit system. For instance, as Table V shows, bus lines that pattern No. 2 covers is double that of pattern No. 6, however, the passengers that they transport is less than those of pattern No. 6. The problem seems worse when we compare pattern No. 5 with pattern No. 10. As the trajectories in pattern No. 5 are about 2.3 times trajectories in pattern No. 10 whereas the bus lines of pattern No. 5 only have less than 10% of the bus lines of pattern No. 10. From this observation, we suggest adding new bus lines in the corresponding area (e.g., patterns No. 5 and No. 6) in order to relieve the traffic pressure and balance relative transportation capacity. Next, we analyze the transportation capacity in a wired pattern. We take pattern No. 1 in Table IV as an example, since this pattern has more trajectories and more bus lines. As Table VI shows, as for bus lines that pattern No. 1 covers, there is a huge divergence on the number of trajectories. It means that some bus lines e.g., No. 667, No. 807 transport more passengers while some lines e.g., No. 666 and No. 648 transfer much less. Therefore, we suggest the Beijing Public Transport Group add buses for the bus lines No. 667 and No. 807 under heavy transportation pressure.

VI. Conclusion

So far, research work on evaluating public transit system by using big data analytics has been very few. From big data analytics perspective, we collect large-scale OD pairs of bus trips in Beijing and cluster the OD pairs by a three-phase clustering strategy ODTC. Based on the clustering results, we compare the relative transportation capacities of some frequent OD pairs, further give some suggestions on bus line adjustment and inter-zonal bus planning. The case study can be viewed as a guideline to reach the practical feasibility of the cluster strategy and high applicability of cluster result analyses.

Acknowledgment

This work is supported by National Key R&D Program of China (No. 2017YFC0803307) and National Natural Science Foundation of China (No. 61472408).

References

[1] R. H. Güting, F. Valdés, and M. L. Damiani, "Symbolic trajectories," *ACM Trans. Spatial Algorithms Syst.*, vol. 1, no. 2, pp. 7:1–7:51, 2015.

[2] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 29:1–29:41, 2015.

[3] G. Yuan, P. Sun, J. Zhao, D. Li, and C. Wang, "A review of moving object trajectory clustering algorithms," *Artif. Intell. Rev.*, vol. 47, no. 1, pp. 123–144, 2017.

[4] A. Mace, R. Sommariva, Z. Fleming, and W. Wang, "Adaptive k-means for clustering air mass trajectories," in *Intelligent Data Engineering and Automated Learning - IDEAL 2011*, 2011.

[5] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: A partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, 2007, pp. 593–604.

[6] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger segmentation using smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1537–1548, 2015.

[7] F. Zhang, B. Jin, T. Ge, Q. Ji, and Y. Cui, "Who are my familiar strangers?: Revealing hidden friend relations and common interests from smart card data," in *Proceedings of the ACM International on Conference on Information and Knowledge Management*, 2016, pp. 619–628.

[8] S. Atev, G. Miller, and N. P. Papanikolopoulos, "Clustering of vehicle trajectories," *Trans. Intell. Transport. Sys.*, vol. 11, no. 3, 2010.

[9] J. Bao, T. He, S. Ruan, Y. Li, and Y. Zheng, "Planning bike lanes based on sharing-bikes' trajectories," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1377–1386.

[10] G.-P. Roh and S.-w. Hwang, "Nncluster: An efficient clustering algorithm for road network trajectories," in *Database Systems for Advanced Applications*, 2010.

[11] W. Hu, X. Li, G. Tian, S. Maybank, and Z. Zhang, "An incremental dpmm-based method for trajectory clustering, modeling, and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1051–1065, 2013.

[12] D. Hallac, S. Vare, S. Boyd, and J. Leskovec, "Toeplitz inverse covariance-based clustering of multivariate time series data," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 215–223.

[13] B. Han, L. Liu, and E. Omiecinski, "Road-network aware trajectory clustering: Integrating locality, flow, and density," *IEEE Transactions on Mobile Computing*, vol. 14, no. 2, pp. 416–429, 2015.

[14] H. Binh, L. Ling, and O. Edward, "A systematic approach to clustering whole trajectories of mobile objects in road networks," *IEEE Trans. on Knowl. and Data Eng.*, vol. 29, no. 5, pp. 936–949, 2017.

[15] Z. Wang, B. Jin, F. Zhang, R. Yang, and Q. Ji, "Discovering trip patterns from incomplete passenger trajectories for inter-zonal bus line planning," in *Network and Parallel Computing*, 2016, pp. 160–171.

[16] T. Caliski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[17] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 952–961.

[18] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 262–272.