Station Function Discovery: Exploring Trip Records in Urban Public Bike-Sharing System

YAN GUO, XINGFA SHEN, QUANBO GE, and Landi Wang

1Department of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, 310018. (e-mail:95297517@qq.com)
2Department of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, 310018. (e-mail: shenxf@hdu.edu.cn)
3Department of Automation, Hangzhou Dianzi University, Hangzhou, China, 310018. (e-mail: qbge@hdu.edu.cn)
4Department of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, 310018. (e-mail: 741510967@qq.com)

Corresponding author: Quanbo Ge (e-mail: qbge@hdu.edu.cn).

This work was supported by the cross-discipline innovation team building project of Hangzhou Dianzi University and the Natural Science Foundation under Grant No.61672198.

ABSTRACT As a new part of public transportation system, urban Bicycle-Sharing System (BSS) consists of bike stations with various kinds of functions, which have significantly important impact on station planning, user pricing, advertisement distribution and so on. After being adopted and deployed in more and more cities, the BSSs accumulate increasingly huge usage data (i.e., trip records) which closely relates to the social and economic activities of users in the city. Therefore, it is possible to take advantage of BSS usage data to infer the functions each station has and then get the functions of regions where the station located. Based on the historical trip records dataset of users, a machine learning algorithm, Latent Dirichlet Allocation (LDA), is adopted to learn the functions of bike stations. Furthermore, k-means clustering algorithm is used to cluster these stations based on their functional profiles. We implement our method using the real-world dataset generated by more than 330 bike stations during three months in Capital Bikeshare system. The proposed station function discovery method is validated by the analysis of spatiotemporal characteristics on traffic patterns for station clusters and evaluated by the comparison of clustering results with the data of point of interests (POI) and station names.

INDEX TERMS Bike-Sharing System, Station Function Discovery, Trip Records, Latent Dirichlet Allocation, K-means Clustering

I. INTRODUCTION

With the sustainable development of the city, public Bike-Sharing System (BSS) plays an increasingly important role in modern urban transportation system construction [1]. BSS has been launched in more than 450 cities around the world, such as Washington, D.C., London and Paris, from 1,000 to 50,000 bikes [2]. Stations in BSS are usually deployed densely and connected seamlessly with bus stops and subway stations in the city, making BSS a perfect option for short-distance travels and an excellent solution for the "last-mile" transit challenge in daily commutes [3] [4].

With the popularization and development of BSS, more and more usage data has been accumulated, such as the user’s mobility which reflected his/her social and economic activities in different locations and time. However, the large-scale and high-dimension characteristics of the historical usage dataset make it challenging to extract features and acquire knowledge from BSS.

Bike stations in BSS usually have different functions (i.e., operational modes) owing to their spatial distribution and demographic characteristics [5]. The functions resemble those of urban regions in the sense that they are both aggregated in space and time in order to fulfill certain purposes of people. It is inaccurate or even infeasible to perform the function discovery of urban regions directly based on BSS data, because BSS users are just a subset of all residents and visitors in the city. However, it’s possible to learn station functions by the usage data of BSS.

We put forward a data-driven station function discovery method by utilizing the usage data (i.e., trip records). We regard each station as a document and station functions as topics of one document by a machine learning algorithm,
LDA (Latent Dirichlet Allocation), which is adopted to discover station functions distribution based on the historical dataset of user trip records. Then stations are clustered by their functions with k-means algorithm.

Taking point of interest (POI) data and station names as ground truths, the proposed method exhibits a quite good performance in clustering accuracy. In addition, the Voronoi diagram could be made to divide the city into numerous regions by a set of seed points (i.e., bike stations).

The core components of the method are LDA and k-means. In addition, we select the dataset of Capital Bikeshare system to evaluate the performance of the proposed method.

Specifically, the major contributions of this work are as follows:

- We present the function (i.e., topic) discovery method by combining LDA model and k-means clustering algorithm. We regard each station as a document, each user behavior at a bike station as a text word and station functions as subjects of one document. The function distribution probability of each station generated by LDA, is utilized to cluster the stations with similar functions by k-means.

- We conduct the analysis of traffic patterns in station clusters with different functions. The Voronoi diagram is utilized to partition the entire BSS areas into a lot of regions based on the spatial information of bike stations, where each polygon represents a functional region covered by a public bike station.

- In order to evaluate the performance of the method, its clustering results are compared with the results clustering by POIs dataset (e.g., restaurants and shopping malls) and station names. After verification, it proves that our method is feasible. Thus, it is possible that using the data of public bicycles only can infer the function of the area, which would help urban planning, corporation site selection, advertisement distribution and etc.

The rest of this paper is organized as follows. The related work is discussed in Section II. Section III and Section IV introduce the main design of the proposed method and its implementation, followed by the spatiotemporal analysis of traffic patterns with discovered functions in Section V. Section VI shows the performance evaluation of the proposed method by using POI data and station names as ground truths. Section VII concludes the paper.

II. RELATED WORK

A. URBAN REGIONAL FUNCTION DISCOVERY

As an important component of urban computing, urban regional function analysis is utilized to learn the land usage of urban residents such as residential region, business region, leisure area.

With the rapid development of sensing technology and computing environment, more and more researchers have studied how to realize the function discovery of urban area through crowd mobility and activity pattern. Most of them choose urban taxi data for urban function discovery. Depending on the qualitative and quantitative analyses of taxi passengers data, Guande Qi et al. [6] concluded that the behavior of taxi passengers can express the social function of the area. They used the real large-scale taxi dataset for experimental analysis. The results showed how to identify the three most typical functional regions of the city by a very simple way. Jing Yuan et al. [7] first used the main road information to divide the city into several regions by the map segmentation algorithm. What he used to analyse the different functions of urban areas and the core of each function is the LDA algorithm, a subject inference model.

B. BIKE SHARING SYSTEMS

Re-allocation: The increasing demand of public bicycles attracts more and more scholars to study this system. Now some scholars put much effort into the bike scheduling optimization strategy research (i.e., re-allocation operation of bicycle), which can be divided into two types. The first type is the static bicycle reposition problem (SBRP) [8], [9] and the second is defined as a dynamic bicycle reallocation problem (DBRP) [10], [11]. Yao et al. [12] proposed a hybrid bicycle allocation strategy for bicycle lifetime optimization, which effectively reduced the degree of imbalance of system load.

Clustering: Other researchers have used this historical usage records. They mainly put focus on the following two directions: clustering and prediction. Adam Wilkinson Davis et al. [13] analyzed the BSS in Northern Ireland and showed that most of public bicycle stations were divided into two types: the commuting mode on weekdays and the leisure mode on weekends.

Predictions: The predictions are also grouped into two categories: model-based and model-free. The model-based predictions include ARIMA, ARMA, Kalman filter and so on. Model-free methods include machine learning, neural networks, non-parametric regression, etc. Christian Lee et al. [14] used three machine models: Poisson distribution, neural network and Markov chain, to predict the availability of Capital Bikeshare system. Yao et al. [15] took into account two situations to infer the demand estimation that users are not served occur in real world.

III. MAIN DESIGN

Fig. 1 shows the main design of the method which consists of four steps. Firstly, the city area is partitioned into sub-regions using the locations of stations as seeds with the Voronoi diagram algorithm. Secondly, the LDA algorithm is adopted to learn the function probability distribution of stations. Thirdly, the k-means clustering algorithm is adopted to cluster the stations based on the topics of stations. Finally, the results of operational mode and discovery method are validated with the TF-IDF algorithm based on POI and station name dataset.

A. LDA-BASED STATION TOPIC DISCOVERING
Each trip of a passenger generates record that constituted the

2) Usage Record Model

Each trip of a passenger generates record that constituted the usage record dataset. Each record is defined as:

\[ T_S = (T_{S,O}, T_{S,D}, T_{S,R}, T_{S,TD}) \]  

(1)

where, \( T_{S,O} \) refers to the rental station (origin station), \( T_{S,D} \) represents the return station (destination station), \( T_{S,R} \) refers to the renting time, \( T_{S,TD} \) refers to the return time. So the OD record consists of its spatial and temporal attributes.

3) User Travel Model

We can get two user travel modes from the usage data of BSS: the rent mode and the return mode. Each model is composed of three elements. We define \( M_O \) as the rent model, and \( M_D \) as the return mode:

\[ M_O = (T_{S,O}, T_{S,D}, T_{S,TD}) \]  

(2)

\[ M_D = (T_{S,O}, T_{S,D}, T_{S,TD}) \]  

(3)

Where, \( M_O \) contains the rental station, return station and renting time of users, and \( M_D \) contains the rental station, return station and return time.

4) Station Usage Model

The usage modes of stations reflect the usage patterns of different users in different periods and the transition relationship with other stations. Similar to the user travel mode, there are two kinds of station usage modes: rent mode and return mode, defined as \( X_{SO}, X_{SD} \), respectively. From the user travel modes, we can derive the usage modes vector for all stations:

\[ X_{SO} = (C_{O1}, C_{O2}, \ldots, C_{O1}, \ldots, C_{Os}) \]  

(4)

\[ X_{SD} = (C_{D1}, C_{D2}, \ldots, C_{D1}, \ldots, C_{Ds}) \]  

(5)

Where, \( C_{Oi} \) is the record of all trips occurred at the station \( s \), which is an \( S \times T \) matrix.

\[ C_{Oi} = \| [M_O = (x,y,z)|x = s, y = i, z = t] \| \]  

(6)

Each entry in \( C_{Oi} \) denotes a trip record that the bike is rented from the station \( s \) at the time point \( t \) and returned to the station \( i \) (\( i = 1 \ldots S \)).

\[ C_{Di} = \| [M_D = (x,y,z)|x = i, y = s, z = t] \| \]  

(7)

Each entry in \( C_{Di} \) denotes a trip record that the bike is rented from the station \( s \) (\( s = 1 \ldots S \)), and returned to the station \( s \) at the time point \( t \) (\( t = 1 \ldots T \)).

5) Modeling of BSS Based on Document-Topic Model

In a document, each word is obtained by the process of "selecting a topic with a certain probability and choosing a word from the topic with a certain probability". Thereby, the station function can be deduced from the user travel modes with LDA.

Assuming that there are a total of \( S \) stations in BSS, the number of the document we have studied is also \( S \). According to the usage mode vectors \( X_{SO} \) and \( X_{SD} \) of stations, we can derive the components of any station which defined as \( C_i = (W_{Oi}, W_{Di}) \).

To illustrate the connection between words and station \( i \), we define \( W_{Oi} = C_{Oi}, W_{Di} = C_{Di} \). The words are generated as shown in Fig. 2, where the X-axis is the time and the Y-axis is the number of the station. At time \( t \), the number of OD records originated from the station \( i \) to the other stations numbered \( 1 \ldots S \) is \( M \), which denotes the station \( i \) have \( M \) words, that is, \( W_{Oi} = (i, s, t) \). Similarly, the number of OD records originated from the station \( 1 \ldots S \) to the station \( i \) is given by \( M \) at time \( t \), so the station \( i \) has \( M \) words like \( W_{Di}(s, i, t) = M \).
We use the LDA model to perform function mining on stations. The definitions of the symbols are listed in Table 1. The generative can be described as follows:

1) For each topic \( k \in [1,k] \), draw \( \varphi_k \sim \text{Dir}(\beta) \).

2) Given the \( s \)-th document \( s \in [1,s] \) in corpus \( D \), draw document-topic distribution \( \theta_{i} \sim \text{Dir}(\alpha_{i}) \), and draw the length of the document \( N_{s} \sim \text{Poiss}(\bar{\alpha}) \).

3) For the \( n \)-th word \( n \in [1,N_{s}] \) in the \( d \)-th document \( w_{s,n} \),
   - a) draw the relationship between the topic and the word \( Z_{s,n} \sim \text{Multi}(\theta_{i}) \); 
   - b) draw the word corresponding to the topic \( w_{s,n} \sim \text{Multi}(\varphi_{s,i}) \).

Where, \( \text{Dir}() \) is the Dirichlet distribution and \( \text{Multi}() \) is the multinomial distribution.

### TABLE 1: LDA - Based Leased Point Location Function Discovery Model Parameter Table

<table>
<thead>
<tr>
<th>Variables</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S )</td>
<td>Total number of documents</td>
</tr>
<tr>
<td>( K )</td>
<td>The number of topics</td>
</tr>
<tr>
<td>( V )</td>
<td>The total number of words</td>
</tr>
<tr>
<td>( D )</td>
<td>The corpus of whole words</td>
</tr>
<tr>
<td>( N_{s} )</td>
<td>The total number of words in a document</td>
</tr>
<tr>
<td>( \theta_{i} )</td>
<td>An ( S \times K ) matrix representing the topic distribution of the first document</td>
</tr>
<tr>
<td>( Z_{s,n} )</td>
<td>Generating the lease point function for the ( n )-th word of the leased point numbered ( s )</td>
</tr>
<tr>
<td>( \varphi_{k} )</td>
<td>A ( K \times V ) matrix representing the word distribution over the topic number ( k )</td>
</tr>
<tr>
<td>( \bar{\alpha} )</td>
<td>The parameter of Dirichlet distribution of the prior distribution of subject distributions for each document</td>
</tr>
<tr>
<td>( \beta )</td>
<td>The parameter of Dirichlet distribution of the prior of word distributions for each subject</td>
</tr>
<tr>
<td>( \bar{w} )</td>
<td>A word that can be observed</td>
</tr>
</tbody>
</table>

#### B. K-MEANS STATION CLUSTERING

We select the k-means algorithm to cluster the samples into \( k \) clusters. The specific algorithm is described as follows:

1) Randomly selects \( k \) clustering centroid points \( \mu_{1}, \mu_{2}, \ldots, \mu_{k} \in R^{d} \).

2) The following procedure is repeated until getting the convergence results
   - a) For each sample \( i \), calculating its class to which it belongs
   - b) For each class \( j \), its centroid point is recalculated

\[
C^{(i)} = \arg\min_{j} \left\| x^{(i)} - \mu_{j} \right\|^2 \\
\mu_{j} := \frac{\sum_{i=1}^{m} 1 \{ C^{(i)} = j \} x^{(i)} }{\sum_{i=1}^{m} 1 \{ C^{(i)} = j \}}
\]

Where, \( k \) is the number of clusters, \( C^{(i)} \) represents one of \( k \) classes which closest to the \( i \)-th sample and the value of \( C^{(i)} \) is between 1 and \( k \). The centroid \( \mu_{j} \) represents our estimation center of the samples belonged to the same class.

#### C. CALINSKI-HARABAZ INDEX

The Calinski-Harabaz index was firstly published in the article “A Dendrite Method for Cluster Analysis”, which is widely used to evaluate the quality of clusters [16]. If the ground truth labels are not known, the Calinski-Harabaz index can be used to evaluate the model. The higher Calinski-Harabaz score relates to a model with better defined clusters. For \( k \) clusters, the Calinski-Harabaz score \( s \) is given as the ratio of the dispersion mean between clusters and the within-cluster dispersion:

\[
s(k) = \frac{(T_{r}(B_{k}))/(T_{r}(W_{k}))(M-k)/(k-1)}{}
\]

where, \( B_{k} \) is the dispersion matrix between groups and \( W_{k} \) is the within-cluster dispersion matrix, \( M \) is the number of points in our data.

Given correspondence between station and document-topic model has already been mentioned above, we perform k-means algorithm clustering based on the characteristics of documents, and divide the stations into the same cluster. We use LDA to discover the hidden themes of each cluster and model them to find the most important topics in each cluster. The document’s characteristic is the probability distribution of the document on each topic. The concrete realization is given below. Because the function of the station is greatly influenced by its geographical location, the functions of the stations near the location are similar which be more easily divided into the same cluster.

#### IV. IMPLEMENTATION

The function and planning of the areas where bike stations located are a comprehensive reflection of the dynamic characteristics of the city. Therefore, it is possible to extract and classify the patterns by the geographical information and the usage record information of public bike stations.

We select BSS in Washington, D.C. as the subject of investigation without loss of generality. All of the data used in this article can be found on Capital Bikeshare system website. And we use the cumulative OD usage records from August to October 2014.

#### A. STATION DATA

Capital Bikeshare system in Washington, D.C has a rapid development in the United States. The details of the station data include station number, latitude, longitude, bicycle capacity and the number of parking slots.

#### B. USAGE RECORD DATA

The user uses a public bike with swiping card, which records his mobility information, such as his rent station number, return station number, use duration, rent time, return time, user type. All of these are stored in the service center server of the system and they also can be found on the official website.
C. POI DATA

The lives of people are closely related to Points of Interest (POI), such as schools, supermarkets, restaurants and hospitals. So POIs are specifically divided into shops, public facilities, institutions and buildings according to their geographical indication function or geographical names with humanistic meanings.

The POI data are obtained from the open public website of Washington, D.C. in the United States (http://opendata.dc.gov/). The original POI data downloaded from the website contains many unwanted attributes, such as telephone, area code, house number and zip code. We remove these redundant attributes after excluding null and abnormal data and keep only four aspects of information finally: name, category, latitude and longitude. Thus, a total of 24,952 POI records left in the end. We classify them into 51 kinds of POI records, which are divided into six categories: "inhabitants/business/shops/culture/services/transportation".

V. RESULTS ANALYSIS

We use the k-means method to cluster the stations and the Calinski-Harabasz (CH) method to determine the optimal number of clusters for the initial parameter of k-means. As mentioned in Section III, higher the Calinski-Harabasz score is, better the clustering effect is. Fig. 3 shows that Calinski-Harabasz score is the highest when \( k = 7 \). Therefore, we conclude that \( k \) is chosen as 7.

Table 2 is the statistical results of the subject discovery. As showed in Fig. 4, all stations are distributed on the map in their clusters.

A conclusion can be drawn from Table 2 that the stations in cluster C3, C4 and C5 are used with the first three higher frequencies, which are both located in the city center. Therefore, a hypothesis can be inferred that stations in cluster C3, C4 and C5 may be the core stations linked with others. The core components are commercial, residential and cultural areas, so it is likely that cluster C3, C4 and C5 are located in the areas with these functions. In contrast, cluster C1, C2 both lie on the edge of the city and their average usage frequencies of stations are both relatively low. Cluster C6 is adjacent to C3 and C4 which have a relatively high total frequency of usage. Furthermore, the stations which located in the same cluster are also closed and this phenomenon could illustrate that the use of BSS stations is affected by their spatial locations.

The purpose of bicycle riders is to fulfill their certain social activities, such as commuting, amusement, dining, etc. Therefore, the mobility patterns could reflect the function of

TABLE 2: LDA Model Site Information Statistics

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Total Stations</th>
<th>Total trips</th>
<th>Trips/Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>44</td>
<td>46507</td>
<td>1056</td>
</tr>
<tr>
<td>C2</td>
<td>72</td>
<td>39011</td>
<td>541</td>
</tr>
<tr>
<td>C3</td>
<td>48</td>
<td>242906</td>
<td>5060</td>
</tr>
<tr>
<td>C4</td>
<td>20</td>
<td>173111</td>
<td>8655</td>
</tr>
<tr>
<td>C5</td>
<td>44</td>
<td>245301</td>
<td>5575</td>
</tr>
<tr>
<td>C6</td>
<td>60</td>
<td>206489</td>
<td>3441</td>
</tr>
<tr>
<td>C7</td>
<td>43</td>
<td>30138</td>
<td>700</td>
</tr>
</tbody>
</table>

FIGURE 3: The Number of \( k \)

FIGURE 4: Station Clustering Results of Capital Bike-Sharing System

FIGURE 5: OD Frequency Statistics
BSS stations. In the following parts of this section, we infer the functions of each station based on the clustering results and the analysis of the usage characteristics of public bicycle users across the 24-hours a day on weekdays and weekends, respectively.

A. CLUSTERING ANALYSIS

1) Overall analysis

Fig. 5 is the statistics graph of the traveling frequencies between the different clusters from C1 to C7. The X-axis represents the renting clusters and the Y-axis denotes the returning. It is found that the traveling flows of each cluster are mainly originated from itself. The reason is that BSS is the capillaries in public transportation system providing the "door to door" services and effectively satisfied the demand for short-distance travel. For clusters C1, C2, and C7, users are more frequently to travel in itself, while users in other clusters are more frequently with each other.

The number of rentals in different clusters shows that C1, C2, and C7 are of low frequency, and C4 does not have double peaks which are different from other clusters from Fig. 6(a). What Fig. 6(b) illustrates is that the Y-axis is the sum of rental frequencies and the X-axis represents the days from Monday to Sunday. Rental frequencies of C1, C2, and C7 from Monday to Sunday are less than those of others. C4 presents an upward trend on Friday and Saturday and a downward trend on Sunday, while C5 shows the opposite trend with C4. It is noteworthy that C3 keeps rising on Friday, Saturday and Sunday.

We named the seven clusters according to their usage characteristics, as showed in Table 3. In the next step, we would analyze the usage characteristics of stations in the seven clusters.

2) C4-Memorials/Park/Museums

We firstly analyze the 24-hour traffic pattern of cluster C4. It presents a single-peak state and the use amounts of weekends are larger than those of workdays. As showed in Fig. 6. It means that cluster C4 is located in the area with a scenic spots. Therefore, C4 is named as "Memorials/Park/Museums".

We analyze the usage characteristics in different registration status for stations in C4. The usage amount of casual (unregistered) users in the cluster is larger than the registered showed in Fig. 7. It means that the function of the stations distributed in C4 is scenic spot. Attributed to the fact that most users in C4 are tourists who rent bicycles for short-distance tours between the various attractions. We come to a conclusion that the bicycle rentals originated from stations in C4 usually ends in the same cluster. Public bicycles have provided great convenience to users in the tourism area.

3) C3-Residence1/C6-Residence2

Fig. 8 presents the 24-hour traffic flow patterns for stations in C3. It has a significant double-peak feature on working days. We know that the frequency of bicycle rentals during the morning rush hours (7am ~ 10am) is higher than that
of the evening (4pm ~ 8pm), whereas the pattern of bicycle returns shows the opposite characteristics. This phenomenon is very consistent with people's daily behaviors like commuting from the residential area to the working area in morning and going home from the working area at night. However, there is no such traffic flow patterns and the activities of users increase significantly at night in the weekend.

Fig. 9 shows the traffic flow patterns of users between C3 and other clusters, which denotes that C3 is strongly related to C4 and C5. Moreover, the traffic pattern between C3 and C5 is very consistent with what we called a "residential area ↔ workspace" pattern. Users rent bicycles from C3 and return them to C5 in the morning rush hours, whereas they rent bicycles from C5 and return them to C3 in the evening.

The traffic pattern of C6 is similar to C3, but for some differences. We name C6 as "residence2 (residence/CBD)" according to its user mode. Fig. 6 is the 24-hour traffic flow pattern of C6 with a clear double-peak feature on weekdays, which is similar to C3.

Fig. 5 shows the traffic flow patterns between C6 and other clusters that we infer C6 relates closely to C5. The traffic pattern between C6 and C5 is consistent with the "residential area ↔ workspace" mode. As showed in Fig. 10, users rent bicycles in the morning rush hours from C6 and return them to C5, whereas users rent bicycles from C5 and return them back in the evening rush hours. There are also partial users rent bicycles in the morning from C3 and return them to C6, whereas in the evening rush hours users rent bicycles from C6 and return them to C3. In summary, C6 is the mixed area of living and working.

4) C5-CBD/Busines

From the above analysis, we can see that C5 belongs to the "workspace" pattern. In addition, we get a double-peak phenomenon on weekdays of C5 from Fig. 11. The frequency of bicycle returns in the morning peak (7am~10am) is higher than that of the evening peak (4pm~8pm), but bicycle rentals obey the opposite patterns. Therefore, we name C5 as "CBD".

5) C1-City outskirts1/C2-City outskirts2/C7-Mixture

From Fig. 4, we learn that C1 and C2 are both distributed in the periphery of the city with different usage characteristics. Hence, C1 and C2 are named as "City outskirts 1" and "City outskirts 2", whose usage characteristics at different time are also analyzed as follows.

From Fig. 5, the users of clusters C1, C2, and C7 are more frequently to travel in itself, while users in other clusters are more frequently with each other. The number of rentals
For example, for an arbitrary station, commercial buildings, car parks, rentals, hotels and commuter types of geographical objects. We analyze the distributions of POI data to reflect the distribution of certain functions where distribute over different clusters from Fig. 12, we conclude that the relationships between C7 and C3/C5 are close. Travel between them also occurs at the highest frequency during the two peaks of the day. Accordingly, we determine C7 as a mixed-functional area, the region with industry, commercial and residential functions mixed together.

We use this feature to guide the bike redistribution strategy of BSS to balance the utilization rate of stations and slow the situation of "no bike can be rented" and "no slot can be used to return bicycles" down. The use of people’s records on public bicycles can be used to infer the function and distribution of urban areas. This method has great benefits. For example, for C4, business can increase the placement of travel advertisements. C5 is a CBD area where companies can consider to select site here. It is also a good choice to set up entertainment and leisure facilities in the surrounding area. C3/C6 are both residence where could get great effect from real estate advertising.

VI. EVALUATION

A. EVALUATION USING POI

The distribution of POI data reflect the distribution of certain types of geographical objects. We analyze the distributions of POIs and find that POI is more widely distributed on commercial buildings, car parks, rentals, hotels and commuter facilities.

There are 331 public bike stations and we obtain the statistics of POI distribution for any station \((1, 2, ..., i, ..., S)\). For example, for an arbitrary station \(S_i\), we obtain the vector \(P_i = (P_1, P_2, ..., P_p, ..., P_n)\), where \(P_p\) is the distribution of POI for the \(i\)-th station.

\[
POI-S_i(P_p) = \frac{n_p}{N_i} \times \log \frac{S}{||S||} \text{ where } POI \in S_i
\]  

Where, \(n_p\) is the number of the \(p\)-th category POI located in the area of the \(i\)-th station, \(N_i\) is the total number of POI located in the area of the \(i\)-th station, \(S\) is the total number of stations and \(||S||\) is the number of the stations where the \(p\)-th category POI appear. So, the TF-IDF value of the \(p\)-th category POI for the \(i\)-th station is the product of the term frequency of the \(p\)-th category POI for the \(i\)-th station and its inverse document frequency. The product of the term frequency of the \(p\)-th category POI for the \(i\)-th station is the division result of the number of the \(p\)-th category POI for the \(i\)-th station and the total number of the \(p\)-th category POI in the \(i\)-th station.

We further compute the \(POI-C_i(P_1, P_2, ..., P_p, ..., P_n)\) between different clusters:

\[
POI-C_i(P_p) = \sum \frac{||S||}{n_c} \text{ where } POI \in C_i
\]  

Where, \(||S||\) is the number of the \(i\)-th station \(S_i\) belongs to the \(C\)-th cluster and \(n_c\) is the number belonged to the \(C\)-th cluster. The division of the sum of the TF-IDF values of the \(p\)-th category POI belonged to cluster \(C\) and the proportion located in cluster \(C\) in the total number of stations. We also understand this formula as the TF-IDF value of the \(p\)-th category POI. In the end, we calculate the TF-IDF value of each POI category for each cluster, and rank the values.

The detailed result data is showed in Table 4, where FD is the POI frequency density of the stations and IR is its internal ranking.

As showed in Table 4, it can be concluded that for C4, the FD value of culture category including scenic spots, parks and museums are the largest value which indicate that most of the stations in C4 are located near the recreation centers and scenic spots, for example, the famous attractions of United States: "White House", "Jefferson Memorial Hall" and "Kennedy Center".

Residential area is the city’s largest and most basic functional area, which is usually adjacent to the planned commercial area. We conclude that the FD values for C3 and C6 represented for "Residents/Stores/Businesses" are higher than other categories with uniform distribution from Table 4. What indicated by FD value of the resident/business POI results are the main difference between C3 and C6 is that C3 is the more mature residential area, and C6 is a mixture of residential and commercial areas.

Observing the geographical distribution of stations in the two clusters, we find that they are located on the outskirts of the city, except for some exceptions. The most widely distributed category of POI in C1 and C2 is apartments, followed by bus stops and subways.

Urban mixed-function areas are defined as multi-purpose functional areas mixed with industrial, commercial and residential functions where distribute office buildings, residential
TABLE 4: POI Frequency Density And Internal Ranking Among Different Clusters

<table>
<thead>
<tr>
<th>POI</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhabitants</td>
<td>6.689</td>
<td>1</td>
<td>4.494</td>
<td>1</td>
<td>5.759</td>
<td>1</td>
<td>0.936</td>
</tr>
<tr>
<td>shops</td>
<td>0</td>
<td>6</td>
<td>4.373</td>
<td>2</td>
<td>2.851</td>
<td>3</td>
<td>6.304</td>
</tr>
<tr>
<td>business</td>
<td>1.254</td>
<td>2</td>
<td>2.898</td>
<td>3</td>
<td>3.07</td>
<td>2</td>
<td>4.035</td>
</tr>
<tr>
<td>culture</td>
<td>1.211</td>
<td>3</td>
<td>2.192</td>
<td>4</td>
<td>0.759</td>
<td>4</td>
<td>7.595</td>
</tr>
<tr>
<td>services</td>
<td>0.257</td>
<td>5</td>
<td>0.71</td>
<td>5</td>
<td>0.332</td>
<td>5</td>
<td>0.772</td>
</tr>
<tr>
<td>transportation</td>
<td>0.474</td>
<td>4</td>
<td>0.061</td>
<td>6</td>
<td>0.014</td>
<td>6</td>
<td>0.285</td>
</tr>
</tbody>
</table>

VII. CONCLUSION

Due to the rapid development of BSS, station function discovery has been becoming increasingly important. OD data on public bicycles usage records are explored to learn and identify the functions of the stations by utilizing LDA model and k-means clustering algorithm. The experiment is carried out by using the historical dataset of Capital Bikeshare system. Furthermore, POI and station name data are used to validate the results. This method has certain practical value in the function identification of the stations in urban BSS. Future work will be based on the results of efficient bicycle redistribution, reasonable pricing and accurate advertising distribution.

REFERENCES

YAN GUO received the B.S. degree in computer science from Anhui University of Technology, China, in 2017. Currently, she is working as a Master student at the School of Computer Science and Technology, Hangzhou Dianzi University. Her current research interests are in intelligent transportation systems and mobile computing.

XINGFA SHEN received the B.S. degree in electrical engineering and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2000 and 2007. Currently, he is working as an Associate Professor at the School of Computer Science and Technology, Hangzhou Dianzi University. His current research interests are in Big Data in Transportation Systems, CPS and mobile computing.

QUANBO GE received the B.S. and M.S. degrees from the College of Computer and Information Engineering, Henan University, Kaifeng, China, and the Ph.D. degree from Shanghai Maritime University, Shanghai, China, in 2002, 2005, and 2008, respectively. Currently, he is working as a Professor at the School of Automation, Hangzhou Dianzi University. His current research interests are in data fusion, wireless sensor networks, etc.

LANDI WANG received the B.S. in computer science from Renai College of Tianjin University, Tianjin, China, in 2014. Currently, she is working as a Master student at the School of Computer Science and Technology, Hangzhou Dianzi University. Her current research interests are in wireless sensor networks and CPS.