IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received September 22, 2017, accepted October 21, 2017, date of publication October 25, 2017,
date of current version December 22, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2766237

IS2Fun: Identification of Subway Station Functions Using Massive Urban Data

JINZHONG WANG[1,2], XIANGJIE KONG[1], (Senior Member, IEEE), AZIZUR RAHIM[1], FENG XIA[1], (Senior Member, IEEE), AMR TOLBA[3,4], AND ZAFER AL-MAKHADMEH[5]

[1]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China
[2]Shenyang Sport University, Shenyang 110102, China
[3]Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia
[4]Mathematics Department, Faculty of Science, Menoufia University, Shebin El-Kom 32511, Egypt
[5]Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia

Corresponding author: Xiangjie Kong (xjkong@ieee.org)

This work was supported by the Deanship of Scientific Research, King Saud University through research group under Grant RGP-1438-27, in part by the National Natural Science Foundation of China under Grant 61572106, in part by the Natural Science Foundation of Liaoning Province, China, under Grant 201602154, in part by the Dalian Science and Technology Planning Project under Grants 2015A11GX015 and Grant 2015R054.

ABSTRACT Urbanization and modernization accelerate the evolution of urban morphology with the formation of different functional regions. To develop a smart city, how to efficiently identify the functional regions is crucial for future urban planning. Differed from the existing works, we mainly focus on how to identify the latent functions of subway stations. In this paper, we propose a semantic framework (IS2Fun) to identify spatio-temporal functions of stations in a city. We apply the semantic model Doc2vec to mine the semantic distribution of subway stations based on human mobility patterns and points of interest (POIs), which sense the dynamic (people's social activities) and static characteristics (POI categories) of each station. We examine the correlation between mobility patterns of commuters and travellers and the spatio-temporal functions of stations. In addition, we develop the POI feature vectors to jointly explore the functions of stations from a perspective of static geographic location. Subsequently, we leverage affinity propagation algorithm to cluster all the stations into ten functional clusters and obtain the latent spatio-temporal functions. We conduct extensive experiments based on the massive urban data, including subway smart card transaction data and POIs to verify that the proposed framework IS2Fun outperforms existing benchmark methods in terms of identifying the functions of subway stations.

INDEX TERMS Urban big data, data analytics, human mobility, points of interest, subway stations.

I. INTRODUCTION

Urbanization and modernization accelerate the evolution of urban morphology and the related studies on smart cities [1], [2]. To enable smart cities, the priority is to identify the functions of urban regions. Urban planning and human mobility patterns can greatly impact the functions of different urban regions [3]. For example, high-tech development regions are generally assigned by urban planners, while flea markets are developed based on inhabitants' lifestyles. Identifying the functions and proximities of different urban regions can help us sense the pulse of our world and broaden the range of valuable applications such as trip planning, advertisement sites selection, and social recommendation.

The rapid development of network communication and sensing technology make it possible to explore the functions of urban regions by leveraging multi-source and heterogeneous massive urban data rather than traditional approaches that are tedious and inefficient. These big data include social media check-in data [4], traffic trajectory data [6]–[9], GPS data [10]–[13], public transit transaction data [14]–[16], and mobile call records [17], which can provide paramount information and patterns about a city via data integration and analysis. It can greatly help improve the life quality of the residents in the city by better design of the urban region functions.

There exist a number of studies on identifying functional regions based on a variety of urban big data. Karlsson *et al.* [20] illustrate a brief view on clustering algorithm and provide a solid foundation for the following related research. Traditional methods leverage high-resolution satellite imagery to measure land development in urban regions [21], but it can not meet the requirements of timeliness and low cost. Zhi *et al.* [22] propose a new low rank approximation based model to identify five significant types of functional clusters. Pu *et al.* [23] employ resilient back propagation algorithm to discover urban functional regions and the accuracy is up to 90%. Assem *et al.* [24] propose a novel clustering approach to detect urban functional regions.

Kraft *et al.* [25] introduce the concept of local minimum of transport intensities to delimit functional regions based on the car transport flows. Fan *et al.* [26] employ an unsupervised feature learning algorithm for land-use scene recognition based on remote sensing data. Furthermore, Yin *et al.* [27] propose a novel method to identify urban boundaries based on human interactions from over 69 million geo-located tweets. Rudinac *et al.* [28] leverage a convolutional neural network to identify functional regions based on social multimedia data and mine the distribution of latent topics from their annotations. In addition, Peng *et al.* [29] focus on the relationship between the change of urban ecological land and its driving force and acquire some valuable regularities. Kong *et al.* [30] take into account the relationship between picking up and dropping off to identify different urban functional regions, and recommend the best locations for taking a taxi.

There also exist a few works on identifying the functions of urban zones based on human mobility patterns. Qi *et al.* [31] leverage the variation of get-on/off amount to measure the social function of a region, and then recognize three types of functional regions. Yuan *et al.* [32] introduce the LDA model to explore the territory of different functional regions based latent activity and location semantics. Arkar *et al.* [33] reveal the latent functional links between zones through constructing functional matrices. Several functional clusters such as residential areas and commercial districts are identified to understand the dynamics of a city more deeply. Furthermore, the outcomes give a valuable reference for future urban planning. The above-mentioned works mainly focus on exploring the functional regions in a city and do not consider the service functions of subway stations. Subway stops are as densely populated venues and are very important nodes in public transit operations. In this work, we mainly focus on discovering the spatio-temporal functional stations in subway network rather than functions of urban regions. We propose the semantic framework (IS2Fun) to identify spatio-temporal functional stations, and show a better performance compared with other existing approaches. This is the first work which introduces the Doc2vec model to mine the latent service functions of subway stations based on human mobility patterns and points of interest (POIs).

We take into account semantic discovery from two folds. First, we consider human mobility patterns by subway, which are correlated with the service function of a station. Human mobility patterns are mainly hidden in people's daily travel trajectories, i.e., trip origin, trip destination, trip time, pick-up time, and drop-off time. Intuitively, the passenger flow by subway from residential districts to commercial and business areas is for working in the morning rush hours, and is off duty from workplaces to home or other entertainment sites in the evening rush hours. In addition, in terms of the characteristics of people travel behaviors, the station with a large passenger flow in the evening or on weekends is likely to reflect the functions of leisure and entertainment.

Second, we take into account the distributions of POIs around stations as a static feature. The POI configurations reflect the service functions of subway stations to some extent. For example, a station is near to a famous place of interest and is very likely to be a scenic spot function. In addition, the similar distributions of POIs in some stations may show different service functions. For example, a well-known university may determine the educational function for a station, but a common college could denote the residential function for a station.

In this paper, we propose a semantic model-based framework named IS2Fun to identify the service functions of subway stations based on the real-world data (subway smart card transaction data and POIs), which are obtained by introducing Doc2vec and integrating human mobility patterns and POI configurations around stations. According to the dynamic semantic distributions and the static POI characteristics of each station, we aggregate subway stations into different functional clusters. Subsequently, we conduct a great deal of theoretical exploration and experimental studies based on massive urban real data in consideration of Shanghai subway network. The analytic results verify that the performance of IS2Fun outperforms other two methods leveraging human mobility patterns by subway and the distributions of POIs around subway stations.

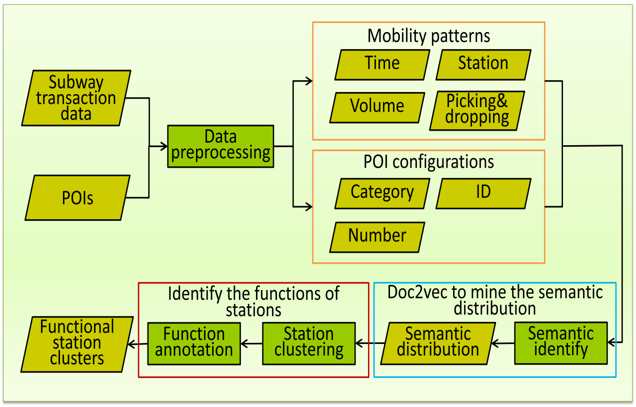


FIGURE 1. The IS2Fun framework for identifying the functional stations.

II. METHODS

A. OVERVIEW

As shown in Fig. 1, we propose a data-driven semantic framework IS2Fun to identify the functional station clusters in a city. First, we preprocess the datasets including the subway transaction data and POIs data from Shanghai in China.[1] We filter outlier data and repeated records with a percentage of 1.5% by data manipulation language. Then, we extract human mobility patterns from four dimensions (time, location, picking&dropping, and volume). Subsequently, we combine the acquired human mobility patterns with POIs, which are fed into the semantic model (Doc2vec). We employ Affinity Propagation (AP) algorithm to aggregate 288 subway stations into different functional clusters. Finally, we annotate each functional cluster based on three ways detailed in II-E2.

B. PRELIMINARY

Definition 1 (Mobility Patterns): A mobility pattern is a quintuple extracted from people's transaction data by subway and is defined as $P = (O_s, D_s, T_{ok}, T_{dk}, N)$, where O_s and D_s denote the origin station and the destination station for a trip respectively, T_{ok} and T_{dk} represent the starting time and the arriving time for a trip with k ranges from 1 to 24, and N denotes the times of a specific trip with the same origin-destination (OD).

In this paper, we obtain the human mobility patterns for every station by projecting transaction data records on its matching station. Especially, we take the patterns on weekdays and weekends into consideration separately based on our daily life.

C. POI FEATURE VECTOR

To identify the spatio-temporal functions of subway stations, we make a statistical analysis on the distributions of POIs which are located within a radius of 500 meters around each station. A POI is a specific location such as hotels, restaurants, and shopping mall, where people may have social activities. Specifically, we represent a POI record with category, name, and geographic coordinate. Later, we calculate the distribution of POIs for each subway station and introduce probability distribution v_{ij} to measure the characteristics of j-th POI category around the subway station i in the following formula.

$$v_{ij} = \frac{\lambda * n_{ij}}{\pi r^2}, \quad i = 1, 2, \ldots, 288; \; j = 1, 2, \ldots, 20, \quad (1)$$

where n_{ij} and r_i denote the number of the j-th POI category and the coverage area (500 meters) around the i-th station respectively, λ is a calibration parameter with a value of 1,000.

The distributions of POI feature reflect the functions of subway stations to some extent. We formulate a POI feature vector $f_i = (v_{i1}, v_{i2}, v_{i3}, \ldots, v_{iK})$ where K is the total number of POI categories.

D. SEMANTIC MODELING

In recent years, semantic models play an important role in mining the hidden semantic structure of documents in a corpus [34]. In this model, distributed word vector representation is proposed which contributes to sentiment analysis, semantic mining and information recommendation, and so on. Specifically, a word of documents in a corpus is mapped to the low-dimensional vector space indirectly resulting in the representation of the hidden semantics. Based on the context of words of a document, we can obtain the semantic distribution, and then extract the topic characteristics.

For subway network, people usually transit between different subway stations for their social activities (working, shopping, and traveling), which show the close relationship between stations and human mobility patterns explicitly, and reflect station's latent function implicitly. Therefore, we take the problem of identifying the latent function in a subway station as the problem of mining the hidden semantics of a document. As shown in Fig. 2, we deem a subway station as a document and a function as a kind of semantics. In addition, we consider the human mobility patterns and POIs for a station as words of a document. In other words, we identify the latent function of subway stations according to its dynamic (mobility patterns) and static (POIs) characteristics.

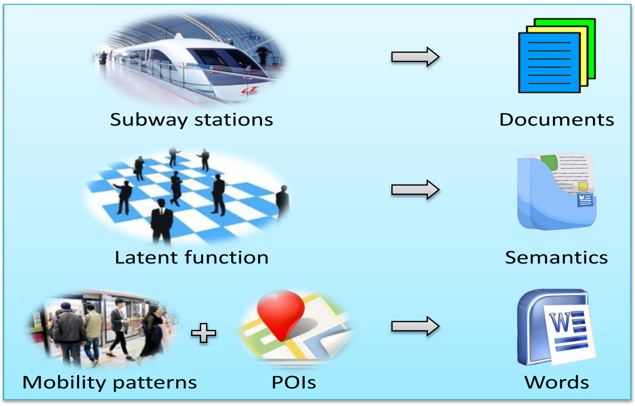
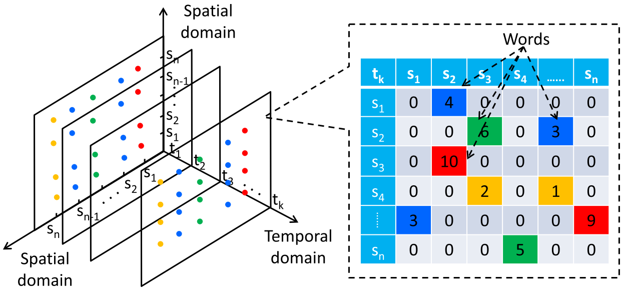


FIGURE 2. Mapping between station function identification and document semantics extraction.



t_k	s_1	s_2	s_3	s_4	......	s_n
s_1	0	4	0	0	0	0
s_2	0	0	6	0	3	0
s_3	0	10	0	0	0	0
s_4	0	0	2	0	1	0
⋮	3	0	0	0	0	9
s_n	0	0	0	5	0	0

FIGURE 3. Mapping between human mobility patterns by subway and words in a corpus.

To describe the analogy more explicitly, we give an example in Fig. 3. First, we preprocess the subway transaction data and clean the dirty data caused by a device fault. Then, we extract the OD trips for each station in every time slot. As defined in Definition 1, we subsequently count the

[1] http://soda.datashanghai.gov.cn/data.html.

mobility patterns related to each station, which are represented as a time series two-dimensional matrix. As shown in the right part of Fig. 3, it shows the mobility patterns (compared with words) in a specific slot, where a nonzero number denotes the intensity of a specific pattern between an origin and a destination.

In this paper, we leverage Doc2vec to identify stations' spatio-temporal functions. Mikolov *et al.* [34] verify that Doc2vec outperforms other semantic model for text classification and sentiment analysis in documents. Doc2vec is an unsupervised model which uses continuous distributed feature vector representations for variable-length fragments of texts (word, sentence, and document). Specifically, Doc2vec represents a document through concatenating paragraph vectors with word vectors in the specific context. Furthermore, the stochastic gradient descent and back propagation are introduced to adjust paragraph vectors and word vectors, in which the former are unique among paragraphs, and the latter are shared.

For each subway station, we extract human mobility patterns and construct an original matrix H of 24×30 elements (24 time slots, 30 days), where each element represents a word in a corpus, and each column represents a paragraph in a document. The mobility patterns' matrix H is as the input and is fed into Doc2vec model, then the output of the model is n-dimensional semantic vector for the station. Given a sequence of mobility patterns $w_1, w_2, w_3, \ldots, w_T$ and column id, the objective of Doc2vec model is to maximize the average log probability as follows:

$$\frac{1}{T}\sum_{t=k}^{T-k} \log(p(w_t|w_{t-k}, \ldots, w_{t+k}, D)). \tag{2}$$

The probability can be modeled using the normalized exponential function (softmax):

$$p(w_t|w_{t-k}, \ldots, w_{t+k}, D) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}, \tag{3}$$

where y_i is un-normalized log-probability for each output pattern i, computed as

$$y = b + Uh(w_{t-k}, \ldots, w_{t+k}; D), \tag{4}$$

where U, b are the softmax parameters. h is constructed by a concatenation of column vectors and pattern vectors extracted from W.

According to distributed vector representation, matrix D consists of columns in matrix H as unique vectors mapping every paragraph in documents, and matrix W consists of elements in matrix H as unique vectors mapping every word. We concatenate the paragraph vectors (column vectors) and word vectors (mobility patterns vectors) to predict the next patterns in a station, which indirectly captures the semantics of station as shown in Fig. 4.

E. FUNCTIONAL STATION IDENTIFICATION

1) STATION CLUSTERING

Based on computing results by Doc2vec model, we cluster subway stations into several different categories in consideration of latent semantic distributions including dynamic mobility patterns and static POI characteristics. By utilizing AP clustering algorithm, stations with the similar function are aggregated into the same cluster. Unlike k-means clustering algorithm, AP does not require to predefine the number of clusters and is able to obtain the better clusters.

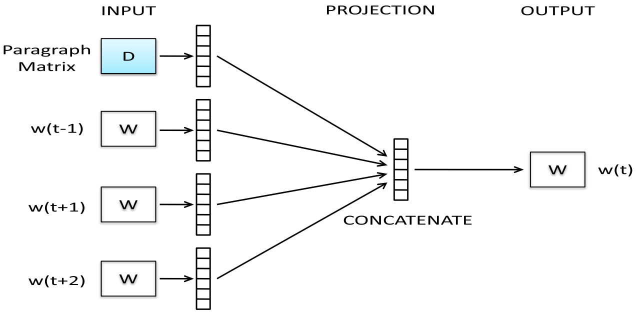


FIGURE 4. Doc2vec model with a distributed memory method.

For station s, we first obtain station's semantic vector F_i by Doc2vec. $F_i = (F_{i,1}, F_{i,2}, F_{i,3}, \ldots, F_{i,n})$ and $F = (F_1, F_2, F_3, \ldots, F_m)^T$, where n and m denote the number of semantics and stations respectively. Then, we calculate and update the matrix R for responsibility and the matrix A for availability iteratively until the algorithm converges. R is generated by data points to candidate centers to indicate how strongly each data point favors the candidate center over other candidate centers, while A is sent from candidate centers to data points to indicate the degree to which each candidate center is available to be a cluster center for the data point. Subsequently, we find the special stations that are representative of clusters and measure the closeness between cluster centers and other stations. Later, all the stations are divided into k clusters on behalf of different spatio-temporal functions.

2) STATION SERVICE FUNCTION ANNOTATION

In terms of stations' semantic distributions and clustering results, we go a step further to annotate stations' service functions. However, there exists a challenging problem on function annotation as a little difference between mining document semantics and identifying station functions. More specifically, for the former, we can leverage the representative words to denote a document topic directly, while it can not be easily fulfilled for the latter (a different application scenario). In other words, semantic vectors generated by latent mobility patterns are not able to name a functional station cluster directly.

In this paper, we leverage the following three aspects to label functional station clusters. First, we use the latent spatio-temporal mobility patterns of each subway station. To our knowledge, people usually travel for their social activities such as working, shopping, and outing, and then generate some hot routes information including origin, destination, time, and volume in subway network. From the prospective of human, the spatio-temporal patterns show stations' spatio-temporal functions in a certain way. For example, one station has more pick-up passenger flow in the morning rush hours and more drop-off passenger flow in the evening rush hours than other stations, which imply that it locates near residential areas and carries commuter passenger flow. Second, we utilize the POI distributions around stations to mark the categories of clusters. We calculate the POI distributions for each cluster and obtain the POI feature vectors f_i shown in Section II-C in detail. In our method, we use TF-IDF algorithm to quantitatively measure the importance level of each POI type both within and between clusters. Finally, we introduce a hand-classified approach to assist the function annotation. Indigenous people may know the functions of a few important stations, e.g. People's Square Station is a station in the commercial district. The aided method contributes to identifying other clusters' spatio-temporal functions.

TABLE 1. **POI category and code.**

POI category	ID	POI category	ID
Government	1	Address information	11
Science & education	2	Food services	12
Motorcycle services	3	Car sales	13
Public utilities	4	Sports & leisure	14
Shopping mall	5	Car services	15
Scenic spot	6	Car repair	16
Finance & Insurance	7	Corporate business	17
Hotel	8	Transport facilities	18
Living services	9	Apartment	19
Health care services	10	Street furniture	20

III. EXPERIMENTS

A. DATASETS

In this paper, we utilize the Shanghai POI dataset in 2015 and subway smart card transaction data in April 2015.

- Points of Interest (POIs): The Shanghai POI dataset mainly contains the category of functional regions, the latitude, and the longitude, which is about 4.9 million records. The dataset represents the static characteristics of different functional subway stations. As shown in Table 1, we show a full list of categories.
- Smart Card Transaction Data of Subway: The dataset contains 451 million transaction records by 14 subway lines and covers the whole month in April 2015, which includes trip origins, trip destinations, trip time, and ticket fare. Through preprocessing the transactions, we obtain the number of passengers for each subway station and the trip times between any two subway stations in different time slots, which show the human mobility patterns by subway.
- Subway Networks: We analyze the subway networks of Shanghai in 2015 and extract subway station names, subway station coordinates, transfer station information, and subway line information. Then, we acquire the statistical results as shown in Table 2.

TABLE 2. **Statistics of smart card transaction data and subway network.**

Data	Name	Result
Card transaction data	Time	April, 2015
	Days	30
	Average trip displacement(km)	11 on weekdays 12 on weekends
	Average trip duration(min)	35 on weekdays 38 on weekends
	Average trip interval(min)	379 on weekdays 264 on weekends
Subway	Subway line	14
	Subway station	288
	Transfer station	47

B. COMPARING METHODS

For identifying stations of different functions, we consider the following methods for comparison.

1) TF-IDF-BASED METHOD

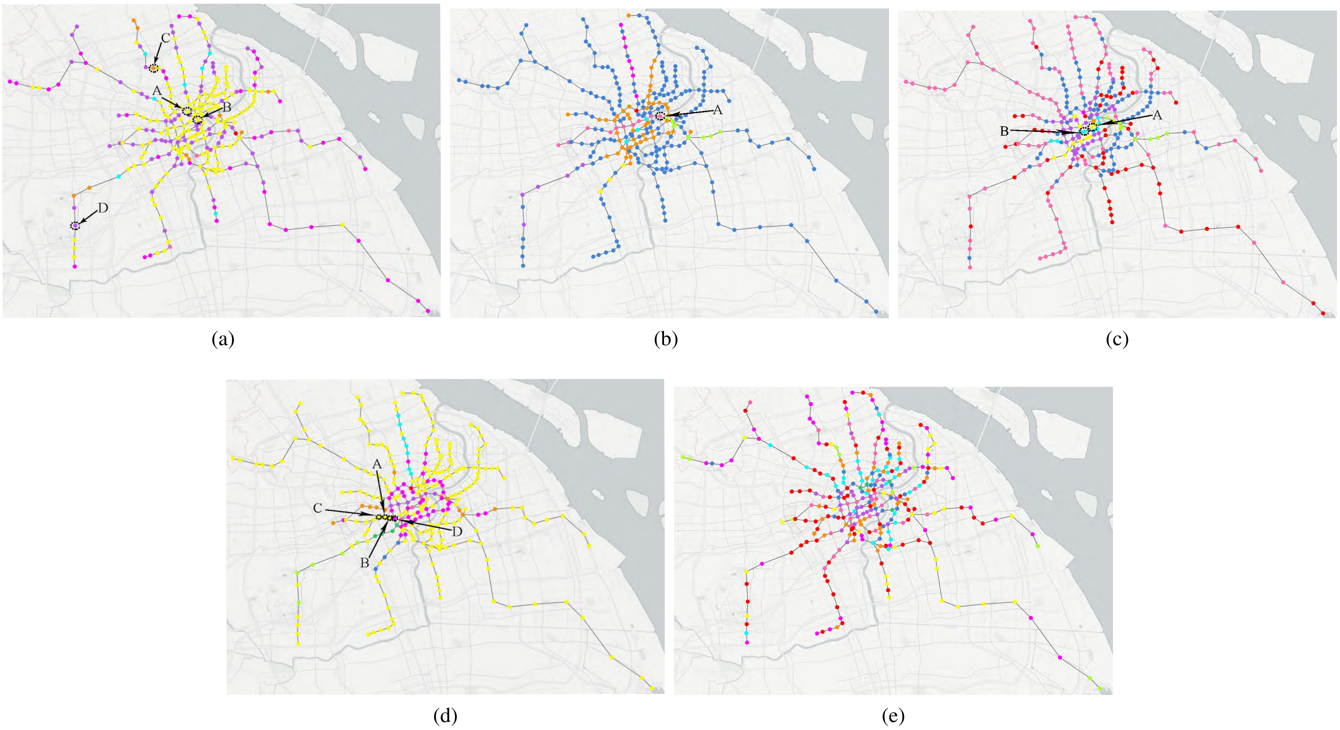
The full name of this method is *term frequency-inverse document frequency* which is one of the most popular term-weighting schemes [35]. We leverage TF-IDF to analyze the distributions of POI feature vectors around the subway stations. Specifically, we formulate a POI feature vector $f_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4}, \ldots, v_{iK})$ within 500 meters from a subway station where v_{ij} is the TF-IDF value of the j-th POI category and K is the number of POI categories. The TF-IDF value v_{ij} is denoted by:

$$v_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} + \log \frac{m}{|s_i : j \in s_i|}, \tag{5}$$

where n_{ij} is the number of the j-th POI category around the i-th subway station. m represents the number of subway stations in Shanghai and $|s_i|$ indicates the number of subway stations that include the j-th POI category. The IDF term is contained by dividing the total number of stations by the number of stations containing the POI category, and then calculating the logarithm of that quotient. Later, we use AP clustering algorithm to cluster the stations into k functional stations.

2) LDA-BASED TOPIC MODEL

LDA-based topic model uses the mobility data by subway in Shanghai. As a probabilistic topic model, it is usually used to mine the latent topic patterns in a corpus [36]. Specifically, we make an analogy between identifying the



(a) (b) (c)

(d) (e)

FIGURE 5. The discovery of functional subway stations based on different methods. Different colours and the same colour of different figures represent different functional cluster types. You may obtain the best viewed in colour. (a) POIs with TF-IDF. (b) Mobility patterns with LDA. (c) Mobility patterns and POIs with LDA. (d) Mobility patterns with IS2Fun. (e) Mobility patterns and POIs with IS2Fun.

functional subway stations and extracting the hidden thematic topic in documents. We regard a subway station as a document, a function as a topic, and human mobility patterns by subway as words in a document. Meanwhile, we use the human mobility patterns as an input of a LDA model. The probability of LDA model is given by:

$$P(W, Z, \theta, \varphi, \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \times \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\varphi_{Z_{j,t}}), \quad (6)$$

where α and β are the parameters of the Dirichlet prior on the per-station topic distributions and the per-function mobility pattern distribution respectively. θ_j denotes the function distribution for subway station m. φ_k represents the mobility pattern distribution of function k. $Z_{i,j}$ is the function for the j-th mobility pattern in subway station i, and $W_{i,j}$ is the specific mobility pattern. Then, we make AP clustering algorithm based on the distribution of mobility patterns and POI distributions for each subway station.

C. EXPERIMENTAL RESULTS

In this section, we analyze experimental results through comparing the proposed IS2Fun with other baselines.

1) IDENTIFICATION OF FUNCTIONAL CLUSTERS

In this paper, we utilize TF-IDF-based approach, LDA-based method, and IS2Fun-based model to identify functional clusters. Furthermore, we consider the characteristics of people's activities and explore the spatio-temporal functions for subway stations on weekdays and weekends respectively. As shown in Fig. 5, TF-IDF obtains 7 functional clusters, while the two other methods acquire 9 and 10 functional classes accordingly.

As shown in Fig. 5a, 7 types of functional clusters are represented by different colors. As the TF-IDF algorithm mainly performs the clustering according to the distribution of POIs around subway stations, the results reflect the static characteristics to some extent and can not focus on the important dynamic factor (human mobility patterns). In particular, Shanghai Railway Station (A) and People's Square Station (B) are aggregated into the same cluster, but their service functions are obviously different illustrated in the other methods. As an important transport hub, Shanghai Railway Station is responsible for handling transit passenger traffic. As for its pervasive connections with the Shanghai street network, the station is also accessible by numerous bus lines and taxi, which contribute to the rush hours for the whole day. However, People's Square Station is surrounded by office buildings, shopping malls, and manifests the commercial service function based on the distributions of POIs. In addition, Shanghai University Station (C) and Songjiang University Town Station (D) are located near universities and should be

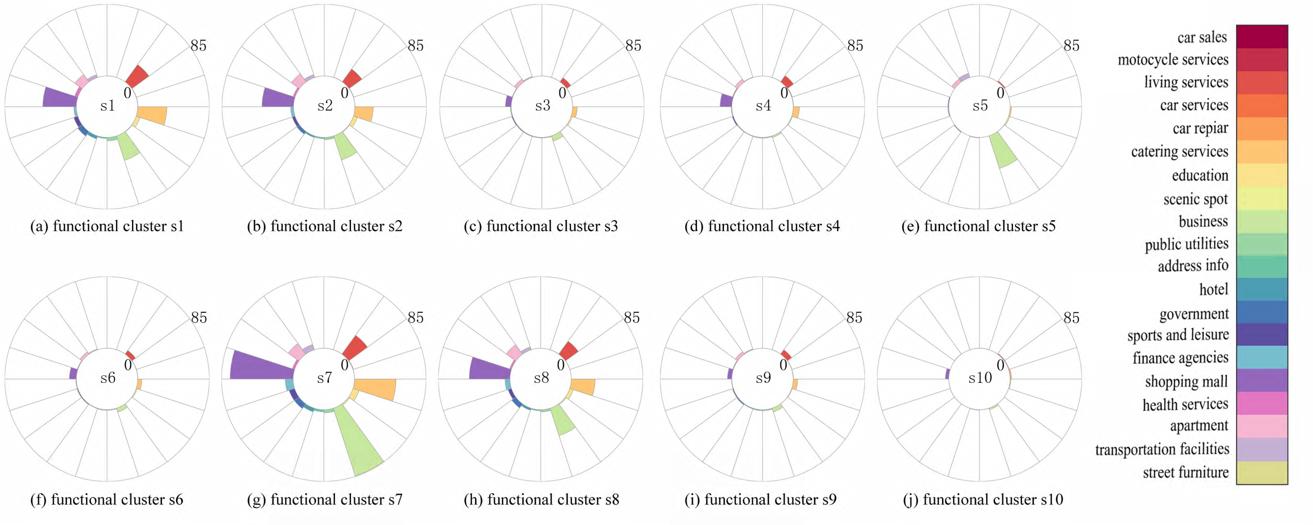


FIGURE 6. The configurations of POIs of different functional clusters $s1$-$s10$.

clustered into one cluster, but the fact that the two stations are divided into two clusters.

In consideration of the characteristics of human mobility, we perform the semantic analysis using LDA-based method. As shown in Figs. 5b and 5c, we obtain 10 functional clusters represented by different colors. Although the clustering results are more reasonable than TF-IDF, there also exist serval stations that are classified into improper functional clusters. For example, Lujiazui Station (A) is located in the middle of the financial district of Lujiazui and belonged to the developed entertainment and commercial regions in Fig. 5b, but it is clustered into the developing commercial and entertainment areas. In addition, as shown in Fig. 5c, South Huangpi Road Station (A) and South Shan-xi Road Station (B) are situated in Huaihai Road CBD and should be classified into the same cluster as the developed business district, but the former is erroneously aggregated into the residential areas. The LDA-based method mainly utilizes the latent topic distributions and overlooks the contexts of human mobility patterns, which contribute to some unreasonable clustering results.

As shown in Figs. 5d and 5e, we leverage IS2Fun to obtain 10 functional clusters. Compared with the LDA approach, the output corrects some functional station categories. But for only using mobility patterns in Fig. 5d, Caohejing Development Zone Station (A) is located in an economic and high-tech industrial development zone, which is wrongly clustered into the old residential regions as the method does not consider the distributions of POIs. Meanwhile, Guilin Road Station (B), Hechuan Road Station (C) and Yishan Road Station (D) are actually all residential areas, which fail to be identified until we feed POI feature vectors into IS2Fun shown in Fig. 5e.

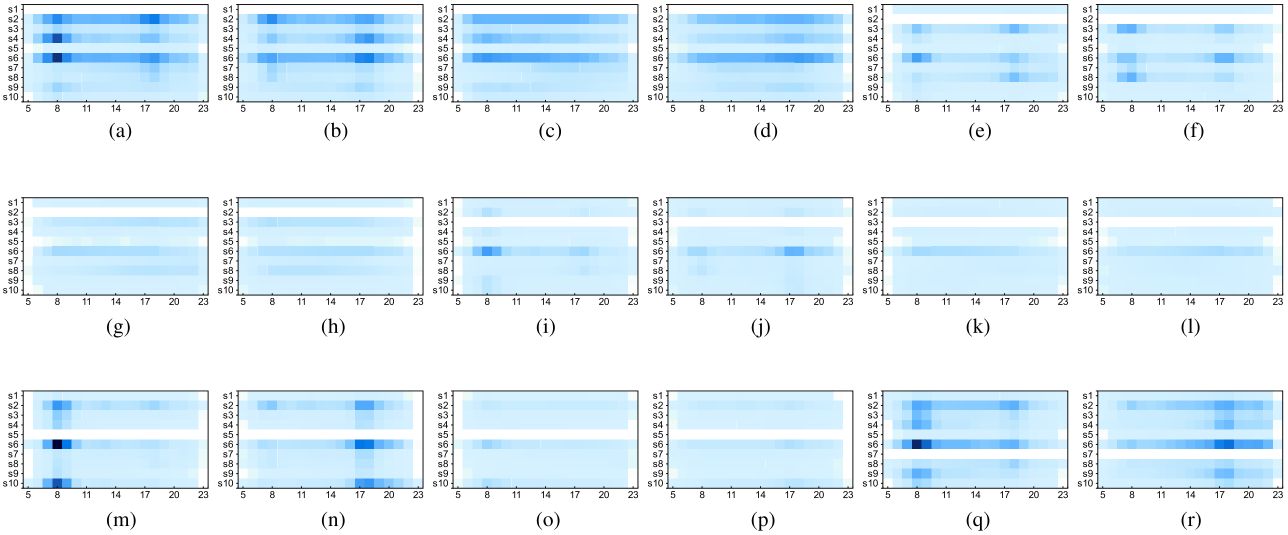
Based on the above-mentioned comparative analysis, we discover that IS2Fun combining POIs and human mobility patterns shows the better performance than other two approaches based on the agreement with labeled functional stations.

2) NOTATION OF FUNCTIONAL CLUSTERS

As shown in Fig. 6, it shows the distributions of POIs density vector of each functional cluster. Each pie slice denotes a POI type with the same color. Furthermore, This plot displays pie slices as lengths extending outward to the edge (0 at inner to 85 at outer). In a single figure, we can notice the ranking of each POI category. Meanwhile, we can also discover the horizontal order of the same POI type. According to my analysis, clusters $s1$, $s2$, $s7$ and $s8$ are more harmonious in urban development and have the high level of urban modernization as opposed to other clusters, as they have more highly ranked POI categories which reflect the characteristics of urban evolution.

a: RAILWAY STATION ($s1$)

The functional cluster has only one station with a distinctive characteristics different from other clusters. As shown in Fig. 6a, there exist more balanced POI configurations which include developed public facilities and top ranking living services (shopping, catering, entertainment, and health care). Combined with human mobility patterns and POI distributions, $s1$ has a large volume of passenger flow for the whole day and has no obvious morning-evening rush hours shown in Figs. 7a - 7d, which is completely unlike commercial and residential clusters. Moreover, we notice that the passenger flow on weekends are less than that on weekdays, but still remain to be massive. Therefore, we annotate the cluster as railway station, that is to say, an urban traffic hub.



(a) (b) (c) (d) (e) (f)
(g) (h) (i) (j) (k) (l)
(m) (n) (o) (p) (q) (r)

FIGURE 7. The spatio-temporal patterns of cluster $s1$, $s2$, $s3$, $s5$, and $s7$ on weekdays and weekends. (a) Arriving, $s1$ on weekdays. (b) Leaving, $s1$ on weekdays. (c) Arriving, $s1$ on weekends. (d) Leaving, $s1$ on weekends. (e) Arriving, $s2$ on weekdays. (f) Leaving, $s2$ on weekdays. (g) Arriving, $s2$ on weekends. (h) Leaving, $s2$ on weekends. (i) Arriving, $s3$ on weekdays. (j) Leaving, $s3$ on weekdays. (k) Arriving, $s3$ on weekends. (l) Leaving, $s3$ on weekends. (m) Arriving, $s5$ on weekdays. (n) Leaving, $s5$ on weekdays. (o) Arriving, $s5$ on weekends. (p) Leaving, $s5$ on weekends. (q) Arriving, $s7$ on weekdays. (r) leaving, $s7$ on weekdays.

b: HIGH-TECH DEVELOPMENT REGIONS ($s5$)

As shown in Fig. 6e, the cluster contains businesses with the first internal ranking and the second external ranking. In addition, transportation facilities rank the second in the horizonal and vertical POI density. The station in the cluster is located in high-tech industrial parks, which verifies the new technology function of $s5$.

As shown in Figs. 7m - 7p, we can discover the arriving and leaving patterns of $s5$ on weekdays and weekends respectively. There exists a huge passenger volume for arriving at 7:00 to 9:00 (Fig. 7m) and leaving at 17:00 to 19:00 (Fig. 7n) during weekdays, while $s5$ has no obvious patterns on weekends. The discoveries coincide with human mobility patterns in working areas and reflect its high-tech development functions.

c: SCIENTIFIC AND EDUCATIONAL REGIONS ($s6$)

The functional cluster contains several representative stations that are close to educational institutions, e.g. Shanghai University, Tongji University, and Shanghai Jiaotong University. Furthermore, the famous electronic market (Qiujiang Road Digital Plaza) lies in the cluster, which exactly reflects the scientific and educational function of $s6$.

d: DEVELOPED COMMERCIAL REGIONS ($s2$)

$s2$ is labeled as the developed commercial regions rooted from our semantic model's outcome. The POI distributions of this functional cluster are similar to cluster $s1$, but in terms of service apartments and financial insurance services. The values in $s2$ are more than that in $s1$. Furthermore, some developed CBDs are located in this cluster.

e: EMERGING COMMERCIAL REGIONS ($s3$)

The functional cluster is annotated as emerging commercial regions since it has an unbalanced POI configuration. In other words, the POI category with high internal ranking suffers from the low external ranking, which reflects the undeveloped condition.

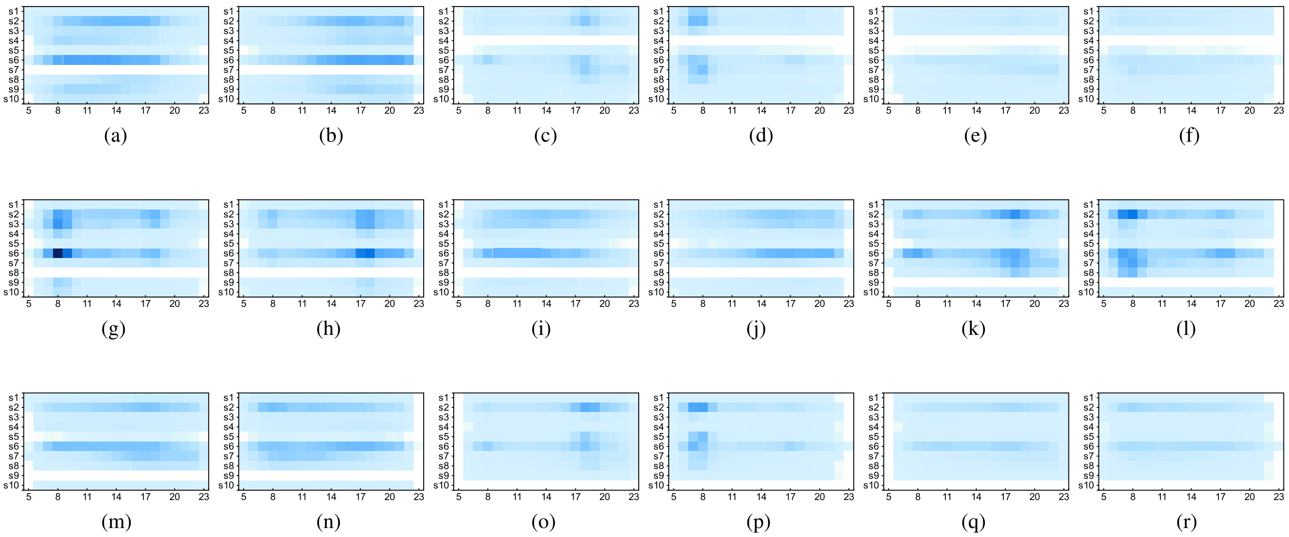
Figs. 7e - 7l show the human mobility patterns in clusters $s2$, $s3$ on weekdays and weekends accordingly. First, we notice that people's activities are more frequent during working days than rest days, which indicates the working venues in the two clusters. Meanwhile, from Fig. 7e and Fig. 7f, people prefer to come to this cluster in the morning rush hours and depart from the areas in the evening rush hours, which conforms to the commercial functions. In addition, in terms of the volume of passenger flow, cluster $s3$ is slightly less than cluster $s2$. Combined with the POI distributions in $s3$, we annotate $s3$ as the emerging commercial regions.

f: EMERGING ENTERTAINMENT REGIONS ($s7$)

In this cluster, we notice that the POI configurations are about the same to cluster $s8$ in terms of POI internal order. However, the POI density of $s7$ is less than that of $s8$. There exist a few food services, shopping services, and entertainment sites, which label the cluster as the emerging entertainment regions.

g: DEVELOPED ENTERTAINMENT REGIONS ($s8$)

The functional areas are some typical leisure and entertainment areas which include some hot shopping mall in Shanghai, such as Westgate Mall, Sogo Store, Cloud Nine Shopping Mall, and Super Grand Mall. Considering human mobility patterns in the cluster, the passenger flow on weekends far outweighs that on weekdays.



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

(m) (n) (o) (p) (q) (r)

FIGURE 8. The spatio-temporal patterns of cluster *s*4, *s*7, *s*8, *s*9, and *s*10 on weekdays and weekends. (a) Arriving, *s*7 on weekends. (b) Leaving, *s*7 on weekends. (c) Arriving, *s*4 on weekdays. (d) Leaving, *s*4 on weekdays. (e) Arriving, *s*4 on weekends. (f) Leaving, *s*4 on weekends. (g) Arriving, *s*8 on weekdays. (h) Leaving, *s*8 on weekdays. (i) Arriving, *s*8 on weekends. (j) Leaving, *s*8 on weekends. (k) Arriving, *s*9 on weekdays. (l) Leaving, *s*9 on weekdays. (m) Arriving, *s*9 on weekends. (n) Leaving, *s*9 on weekends. (o) Arriving, *s*10 on weekdays. (p) Leaving, *s*10 on weekdays. (q) Arriving, *s*10 on weekends. (r) Leaving, *s*10 on weekends.

Intuitively, with the rapid progress of urbanization, many regions have multiple social functions rather than a single function, e.g. Wanda Plaza with commercial and entertainment functions. But we consider the dominant function with the evolution of the spatial and temporal dimension. As shown in Figs. 7q, 7r, 8g, and 8h, there exists a large passenger flow leaving at 20:00 to 22:00 and coming about 18:00, which is mightily for the leisure and relaxing time and is quite different from commercial regions $s2$ and $s3$. As shown in Figs. 8a, 8b, 8i, and 8j, people usually keep active state from 9:00 to 22:00 on weekends, which validate $s7$ and $s8$ functional characteristics more effectively. Meanwhile, the passenger flow and POI configurations of $s8$ are more better than that of $s7$. Therefore, we label the clusters as emerging and developed entertainment areas respectively.

h: OLD RESIDENTIAL REGIONS (s4)

As shown in Fig. 6d, the living services have the highest ranking in the internal order, but its POI configuration is less developed than $s9$. Moreover, there exist some residential buildings and old street districts as opposed to the developed residential regions. So the functional regions are urgent to be developed by administrative departments.

i: DEVELOPED RESIDENTIAL REGIONS (s9)

The cluster is obviously a developed residential areas with the most living services, food services, health care services, and shopping services. In $s9$, people can engage in their daily social activities more conveniently under the sophisticated POI configurations, which contribute to the rapid urbanization process and city planning.

j: EMERGING RESIDENTIAL REGIONS (s10)

We annotate the cluster as the emerging residential areas which attribute to the high ranking of living services in all the POI categories. But the proportion of living services in $s10$ is between $s4$ and $s9$. Although there exist shopping malls, restaurants, hospitals, and banking, they still remain to be further developed.

As shown in Figs. 8c - 8f and Figs. 8k - 8r, we can notice some unique mobility patterns which are different from commercial and entertainment regions. More specifically, people often come to the regions $s4$, $s9$, and $s10$ at 18:00 and go away from 7:00 to 8:00 on weekdays, which are consistent with our daily experiences. Conversely, there do not exist the mobility patterns that set out early and return lately on weekends. Furthermore, more people prefer to leave for the clusters $s2$ and $s7$ on weekends in Fig. 8m and Fig. 8n, which just reflects their commercial and entertainment functions respectively.

IV. CONCLUSION

In this paper, we propose a semantic framework named IS2Fun to identify the spatio-temporal functional station clusters in a city based on human mobility patterns (dynamic characteristics) and POIs (static characteristics). We leverage the massive urban real datasets which include subway transaction records for a whole month and POI information from Shanghai in China in 2015. In terms of mobility patterns, we leverage Doc2vec model to mine the latent travel routes from a view point of spatial and temporal dimension. Combined with POI feature vectors, we obtain 10 functional station clusters, e.g. railway station, educational regions, business regions, and hi-tech development regions.

The experimental results verify that IS2Fun outperforms LDA-based method and TF-IDF-based method based on solely and collaboratively using POIs and human mobility patterns. IS2Fun provides a valuable reference to develop a smart city and improve the operational efficiency of urban public transport.

REFERENCES

[1] P. Sotres, J. R. Santana, L. Sánchez, J. Lanza, and L. Muñoz, "Practical lessons from the deployment and management of a smart city Internet-of-Things infrastructure: The smartsantander testbed case," *IEEE Access*, vol. 5, pp. 14309–14322, 2017.

[2] Z. Ning, F. Xia, N. Ullah, X. Kong, and X. Hu, "Vehicular social networks: Enabling smart mobility," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 16–55, May 2017.

[3] F. Xia, J. Wang, X. Kong, Z. Wang, J. Li, and C. Liu, "Exploring human mobility patterns in urban scenarios: A trajectory data perspective," *IEEE Commun. Mag.*, to be published.

[4] D. Wu, L. Lambrinos, T. Przepiorka, and J. A. McCann, "Facilitating mobile access to social media content on urban underground metro systems," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2016, pp. 921–926.

[5] Q. Zhang, L. T. Yang, X. Liu, Z. Chen, and P. Li, "A tucker deep computation model for mobile multimedia feature learning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 3s, pp. 39:1–39:18, 2017.

[6] X. Kong, Z. Xu, G. Shen, J. Wang, Q. Yang, and B. Zhang, "Urban traffic congestion estimation and prediction based on floating car trajectory data," *Future Generat. Comput. Syst.*, vol. 61, pp. 97–107, Aug. 2016.

[7] X. Yang, A. Chen, B. Ning, and T. Tang, "Measuring route diversity for urban rail transit networks: A case study of the Beijing metro network," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 2, pp. 259–268, Feb. 2017.

[8] P. Zhao and S. Li, "Bicycle-metro integration in a growing city: The determinants of cycling as a transfer mode in metro station areas in Beijing," *Transp. Res. A, Policy Pract.*, vol. 99, pp. 46–60, May 2017.

[9] X. Kong, X. Song, F. Xia, H. Guo, J. Wang, and A. Tolba, "LoTAD: Long-term traffic anomaly detection based on crowdsourced bus trajectory data," in *World Wide Web*. 2017, pp. 1–23.

[10] M. Ni, Q. He, and J. Gao, "Forecasting the subway passenger flow under event occurrences with social media," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1623–1632, Jun. 2016.

[11] F. Zhang, J. Zhao, C. Tian, C. Xu, X. Liu, and L. Rao, "Spatiotemporal segmentation of metro trips using smart card data," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1137–1149, Mar. 2016.

[12] C. Zhong *et al.*, "Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data," *PloS ONE*, vol. 11, no. 2, p. e0149222, 2016.

[13] B. Du, C. Liu, W. Zhou, Z. Hou, and H. Xiong, "Catch me if you can: Detecting pickpocket suspects from large-scale transit records," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 87–96.

[14] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger segmentation using smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1537–1548, Jun. 2015.

[15] J. Zhao *et al.*, "Estimation of passenger route choice pattern using smart card data for complex metro systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 790–801, Apr. 2017.

[16] M. Itoh, D. Yokoyama, M. Toyoda, Y. Tomita, S. Kawamura, and M. Kitsuregawa, "Visual exploration of changes in passenger flows and tweets on mega-city metro network," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 85–99, Mar. 2016.

[17] Y. Chen and C. Shen, "Performance analysis of smartphone-sensor behavior for human activity recognition," *IEEE Access*, vol. 5, pp. 3095–3110, 2017.

[18] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "An improved deep computation model based on canonical polyadic decomposition," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2017.2701797.

[19] Q. Zhang, L. T. Yang, Z. Chen, P. Li, and M. J. Deen, "Privacy-preserving double-projection deep computation model with crowdsourcing on cloud for big data feature learning," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2017.2732735.

[20] C. Karlsson, "Clusters, functional regions and cluster policies," *J. Int. Bus. Studies*, vol. 84, pp. 1010–1018, Mar. 2007.

[21] C. Unsalan, "Measuring land development in urban regions using graph theoretical and conditional statistical features," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3989–3999, Dec. 2007.

[22] Y. Zhi *et al.*, "Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data," *Geo-Spatial Inf. Sci.*, vol. 19, no. 2, pp. 94–105, 2016.

[23] Y. Pu, X. Song, and Y. Ge, "An improved framework to discover functional urban regions using pedestrian trajectories," in *Proc. 24th Int. Conf. Geoinformat.*, Aug. 2016, pp. 1–5.

[24] H. Assem, L. Xu, T. S. Buda, and D. O'Sullivan, "Spatio-temporal clustering approach for detecting functional regions in cities," in *Proc. IEEE 28th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2016, pp. 370–377.

[25] S. Kraft and M. Marada, "Delimitation of functional transport regions: Understanding the transport flows patterns at the micro-regional level," *Geografiska Ann., Ser. B, Human Geograph.*, vol. 99, no. 1, pp. 79–93, 2017.

[26] J. Fan, T. Chen, and S. Lu, "Unsupervised feature learning for land-use scene recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2250–2261, Apr. 2017.

[27] J. Yin, A. Soliman, D. Yin, and S. Wang, "Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data," *Int. J. Geograph. Inf. Sci.*, vol. 31, no. 7, pp. 1293–1313, 2017.

[28] S. Rudinac, J. Zahálka, and M. Worring, "Discovering geographic regions in the city using social multimedia and open data," in *Proc. Int. Conf. Multimedia Modeling*, 2017, pp. 148–159.

[29] J. Peng, M. Zhao, X. Guo, Y. Pan, and Y. Liu, "Spatial-temporal dynamics and associated driving forces of urban ecological land: A case study in Shenzhen City, China," *Habitat Int.*, vol. 60, pp. 81–90, Feb. 2017.

[30] X. Kong, F. Xia, J. Wang, A. Rahim, and S. K. Das, "Time-location-relationship combined service recommendation based on taxi trajectory data," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1202–1212, Jun. 2017.

[31] G. Qi, X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang, "Measuring social functions of city regions from large-scale taxi behaviors," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PERCOM Workshops)*, Mar. 2011, pp. 384–388.

[32] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering urban functional zones using latent activity trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 712–725, Mar. 2015.

[33] S. Sarkar *et al.*, "Effective urban structure inference from traffic flow dynamics," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 181–193, Jan. 2017.

[34] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[35] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 1, pp. 30–44, Jan. 2012.

[36] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

JINZHONG WANG received the B.Sc. degree in computer education from Anshan Normal University, Anshan, China, in 2002, and the M.Sc. degree in computer application technology from Liaoning University, Shenyang, China, in 2005. He is currently pursuing the Ph.D. degree with the School of Software, Dalian University of Technology, Dalian, China. Since 2005, he has been with Shenyang Sport University, Shenyang, China. His research interests include computational social network, network science, data science, and mobile social networks.

XIANGJIE KONG (M'13–SM'17) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently an Associate Professor with the School of Software, Dalian University of Technology, China. He has authored over 50 scientific papers in international journals and conferences (with over 30 indexed by ISI SCIE). His research interests include big traffic data, mobile computing, and cyber-physical systems. He is a Senior Member the CCF and a member of the ACM. He has served as a (guest) editor for several international journals, the workshop chair or a PC member of a number of conferences.

AZIZUR RAHIM received the B.Sc. degree from the University of Engineering and Technology, Peshawar, and the M.Sc. degree in electrical engineering from COMSATS, Islamabad. He is currently a Ph.D. Scholar with the Alpha Lab, School of Software, Dalian University of Technology, Dalian, China, under the supervision of Prof. F. Xia. His research interests include mobile and social computing, ad hoc networks, VANETs, mobile social networks, and vehicular social networks.

FENG XIA (M'07–SM'12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He was a Research Fellow with the Queensland University of Technology, Australia. He is currently a Full Professor with the School of Software, Dalian University of Technology, China. He has authored two books and over 200 scientific papers in international journals and conferences. His research interests include computational social science, network science, data science, and mobile social networks. He is a Senior Member of the ACM, and a member of the AAAS. He serves as the general chair, the PC chair, the workshop chair, or the publicity chair for a number of conferences. He is an (guest) editor of several international journals.

AMR TOLBA received the M.Sc. and Ph.D. degrees from the Faculty of Science, Menoufia University, Egypt, in 2002 and 2006, respectively. He is currently an Associate Professor with the Faculty of Science, Menoufia University. He is currently on leave from Menoufia Univesity to the Computer Science Department, Community College, King Saud University, Saudi Arabia. He has authored/co-authored over 30 scientic papers in international journals and conference proceedings. His main research interests include socially-aware network, Internet of Things, intelligent systems, big data, recommender systems, and cloud computing. He serves as a technical program committee member in several conferences.

ZAFER AL-MAKHADMEH received the M.Sc. and Ph.D. degrees from the Department of Computer Engineering, Faculty of Information and Computer Engineering, Kharkov National Technical University of Ukraine, in 1998 and 2001, respectively. He is currently an Assistant Professor with the Computer Science Department, Community College, King Saud University, Saudi Arabia. His main research interests include cloud computing, social network analysis, big data, and intelligent systems.