Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Discovering Functional Regions in Modern Cities by Using User Check-in Records and POIs

Changyi Ma, Wenye Li
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
changyima@link.cuhk.edu.cn, wyli@cuhk.edu.cn

***Abstract*— The rapid development of the modern city necessitates the division of different functional regions, such as educational regions and business regions. In this paper, we propose an LDA based topic model to discover regions with different functions in a city using user check-in records and Point-of-Interests (POIs). We infer the functions of each region using the proposed model, which regards a region as a document, a function as a topic, categories of POIs and check-in frequency as words. As a result, a region is represented by a distribution of functions, and a function is represented by the distribution of POI categories. We further identify the intensity to illustrate that our method considers both the spatial distance and check-in frequency. We evaluate our method by using a large scale dataset, consisting of two user check-in records that occurred in Los Angeles and Lower Manhattan. The experiment results justify the advantages of our proposed method over the baseline method solely using POIs or check-in records.**

***Index Terms*— LDA, User Check-in Records, POIs**

I. Introduction

The development of a modern city generates the functional regions, which are divided according to the city development plan or formulated by users' daily life. Also, functions may be changed with the development of the city. In this work, we propose a LDA based topic model to discover the functional regions by using user check-in records and Point-of-Interests (POIs). Specifically, we aim to discover the internal thematic meaning of each functional region.

The motivations of our work are as follows. Firstly, the map with different functional regions gives the newcomers a brief understanding of the target modern city. When the user comes to one new city, he/she can find the scenery regions or find some regions he/she prefer according to the discovered functional regions. Secondly, it benefits city planning. For example, if the developer plan to add some shopping malls or markets, the residential region or the educational region should be a good choice. Thirdly, functional region discovery also benefits the places recommendation. When the users in the educational regions, the recommender system will recommend some food places to the users according to their personal preferences.

Corresponding author: Wenye Li (+86-755-84273853).

In this work, the main differences with the existing works are as follows: Firstly, we propose an LDA based topic model [2] to discover the functional regions in the modern city. It is worth to notice that the functional regions are clustered automatically, which is different from the existing works. Secondly, we apply the *Topic Coherence* [9], *Silhouette value* and *Kernel Density Estimation* to evaluate the thematic coherence within each region and also the overall spatial coherence among the functional regions. Besides, we conduct a series of case studies via adopting *POI frequency* [18] to give a brief description of each region. Thirdly, we evaluated the proposed method on a real-world dataset, containing both the semantic information and spatial information.

II. Related Work

Urban Computing[1] The object of Urban computing is to understand human behavior in modern cities via collecting the users' daily records and ubiquitous data, and to provide tools or insights that enhance the real-world applications. There are many existing works [5], [6] studied the taxi drivers' earning by analyzing the getting on and getting off records of the taxis at different locations. Miller et al.[8] analyzed the live vehicle data to find the fastest and shortest route in the target city.

Discovering Functional Regions Discovering Functional Regions [1] is a sub-field of urban planning. Karlsson et al. [7] studied on a series of clustering algorithms using in the field of discovering functional regions. Ge et al.[13] proposed to discover functional regions using the remote-sensor data. Wang et al. [16] analysed the intensity and the move direction over different regions via using taxi trajectories.

Place Recommendation Place recommendation aims to divide the whole city into several small regions to avoid the data sparsity problem. Wang et al. [15] proposed to recommend a set of places to the target users using a spatial pyramid to divide the whole city into several regions with the same acreage. Qi et al. [10] predicted the users' daily move direction by using the getting on/off the amount of taxi passengers. Chen et al. [4] predicted the target users' move direction across regions, and further recommend places

[1]https://en.wikipedia.org/wiki/Urbancomputing

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

according to the time and spatial information within each region.

III. Methodology

A. Functional Region Identification

1) Functional Region Discovery: The input of our model is user check-in records and POIs, formulating the observed variables in the graphical model. Topic z and region r are both the latent variables, word w, POI v, and POI coordinator l_v are the observed variables. N, K and R represent the number of users, topics, and regions, respectively. Fig.1 is the graphical representation of the proposed model. From

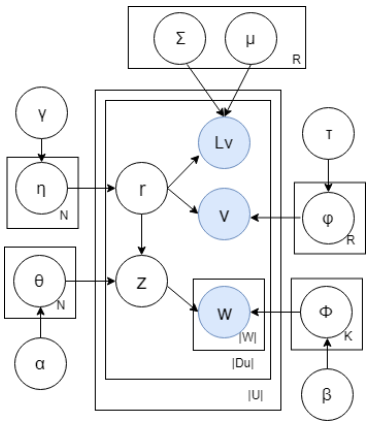


Fig. 1: Graphical Representation of Proposed Model

Fig.1, we observe that the latent variable region v generates POIs v and POIs' coordinators l_v, which means the proposed model considers both density and spatial distance within each region. Topic z generates POIs' categorical words w, which means the topic is correlated with the POIs' category. Here, we regard the topic as the function. Meanwhile, regions r generating topic z means each functional region has a series of topic/functions with different probabilities.

Technically, each topic z is associated with a multinomial distribution over words w. Since most of the humans active around some centers in limited regions. Intuitively, we adopt the Gaussian distribution for each region r, the coordinator of each POI v is characterized by $l_v \sim N(\mu_r, \Sigma_r)$, where the μ_r and Σ_r denote the mean vector and co-variance matrix, respectively. Each region r is associated with a multinomial distribution over POI v.

Priors over η_u, ϕ_z, φ_r are imposed with parameters γ, β, and τ, respectively. We formulate the probabilistic generative process of the proposed model as Algorithm 1.

For the inference process, we apply the mixture of both the EM algorithm and Gibbs sampling algorithm to estimate the parameters. Adopting the existed work [17], we fix the values for the hyper-parameters α = 50/K, γ = 50/R and β = τ = 0.01. After getting the values of hyperparameters, we first initialize the regions using the K-means algorithm, then

Algorithm 1 Generative Process

for each region r **do**
  draw $\varphi_r \sim Dir(\cdot \mid \eta)$;
end for
for each topic z **do**
  draw $\phi_z \sim Dir(\cdot \mid \beta)$;
end for;
for each user u **do**
  draw $\theta_u \sim Dir(\cdot \mid \alpha)$, $\eta_u \sim Dir(\cdot \mid \gamma)$;
end for
for each user check-in records $(u, v, l_v, W_v, t) \in D_u$ **do**
  draw a region index $r \sim Mult(\eta_u)$;
  draw a POI index $v \sim Mult(\varphi_r)$;
  draw a location index $l_v \sim N(\mu_r, \Sigma_r)$;
  draw a topic index $z \sim Mult(\theta_u)$;
  For each word $w \in W_v$, draw $w \sim Mult(\phi_z)$;
end for

randomly initialize the topic assignments. Secondly, updating the region and topic assignment according to the posterior distribution based on the (1), (2).

$$P(r|r_{-u,v}, z, v, l_v, W_v, u, \cdot) \propto \frac{n_{u,r}^{-u,v} + \gamma}{\Sigma_{r'}(n_{u,r'}^{-u,v}) + \gamma} \times \frac{n_{r,v}^{-u,v} + \tau}{\Sigma_{v'}(n_{r,v'}^{-u,v}) + \tau}(l_v|\mu_r, \Sigma_r) \tag{1}$$

$$P(z|z_{-u,v}, r, v, l_v, W_v, u, \cdot) \propto \frac{n_{u,z}^{-u,v} + \alpha}{\Sigma_{z'}(n_{u,z'}^{-u,v}) + \alpha} \times \Pi_{w \in W_v} \frac{n_{z,w}^{-u,v} + \beta}{\Sigma_{w'}(n_{z,w'}^{-u,v}) + \beta} \tag{2}$$

After each iteration, updating the Gaussian distribution parameters based on (3).

$$\begin{aligned} \mu_r &= \frac{1}{|S_r|}\Sigma_{u \in S_r} l_v \\ \Sigma_r &= \frac{1}{|S_r| - 1}\Sigma_{u \in S_r}(l_v - \mu_r)(l_v - \mu_r)^T \end{aligned} \tag{3}$$

The iteration is repeated until convergence. Finally, we calculate the estimated parameters based on (4) respectively.

$$\begin{aligned} &\eta_{u,r}^{sum} += \frac{n_{u,r}^{-u,v} + \gamma}{\Sigma_{r'} n_{u,r'}^{-u,v} + \gamma}, \theta_{u,z}^{sum} += \frac{n_{u,z}^{-u,v} + \alpha}{\Sigma_{z'} n_{u,z'}^{-u,v} + \alpha} \\ &\varphi_{r,v}^{sum} += \frac{n_{r,v}^{-u,v} + \tau}{\Sigma_{v'} n_{r,v'}^{-u,v} + \tau}, \phi_{z,w}^{sum} += \frac{n_{z,w}^{-u,v} + \beta}{\Sigma_{w'} n_{z,w'}^{-u,v} + \beta} \\ &\mu_r^{sum} += \mu_r, \Sigma_r^{sum} += \Sigma_r \end{aligned} \tag{4}$$

2) Evaluation Metrics of Functional Regions: Since the proposed method consider both the semantic information and spatial information, we evaluate the performance from two sides, including topic coherence and spatial coherence.

- **Semantic Part - Topic Coherence** Topic coherence [9] is a common method to evaluate the topic model, in another word, topic coherence reflects whether the single

topic is meaningful to some extent. In this work, discovering the functional regions means the individual region should be topic internal unified and different to other regions. We need to annotate each functional region by summarizing the categorical words, so whether the correlated words are unified is the criterion for the topic evaluation. The existing methods about topic coherence are the extrinsic measure and intrinsic measure, both based on the same high level idea. In this paper, we apply the extrinsic measure, using the pairwise score function Point-wise Mutual Information (PMI) in (5)

$$Score(w_i, w_j) = log\frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{5}$$

where $p(w_i)$ represents the probability of seeing w_i in a document(region), and $p(w_i, w_j)$ means the probability of seeing both w_i and w_j co-occurring in a document. For example, if word *university* appears 10 times in the region, and *college* appears 15 times in the region, and *(university, college)* appears 10 times in the region. The PMI score is log(1/15). Besides, if word *university* appears 10 times and word *coffee shop* appears 2 times, and *(university, coffee shop)* appears 1 time, the PMI score is log(1/200). Based on the above knowledge, we can conclude that the more words with the same category index (information from Foursquare), the more coherence the topic is. In this case, the larger value of topic coherence represents a better result.

- **Spatial Part - Kernel Density Estimation** We estimate the intensity within each region, since the functional regions are not distributed uniformly. Intuitively, user check-in records reflect the popularity of an individual functional region. As a result, we regard the POIs in a functional region as the input of the Kernel Density Estimation model to infer the functional intensity in a specific functional region.
  Suppose there are n POIs $v_1, v_2, ..., v_n$ with coordinates are $x_1, x_2, ..., x_n$, we estimate the intensity at each center of the functional region c, defined as (6):

$$\lambda(c) = \Sigma_{i=1}^{n} \frac{1}{nb^2} K(\frac{d_{i,c}}{b}) \tag{6}$$

  $d_{i,c}$ represents the distance between x_i to c, $K(\cdot)$ is the kernel function whose value decreases with the increasing of $d_{i,c}$. For the kenerl function, we choose the Gaussian function as the kernel functional as (7), i.e,

$$K(\frac{d_{i,c}}{b}) = \frac{1}{\sqrt{2\pi}} exp(-\frac{d_{i,c}^2}{2b^2}) \tag{7}$$

  b is the bandwidth, which determines the smoothness of the estimated density. A larger b leading smoother estimation while a smaller b reflecting more detailed fluctuation. MISE [14] provides an empirical setting criterion, and we fix b as 0.9 in this work.

- **Spatial Part - Silhouette Value** Silhouette value [12] is a measure of how close an object is to its own cluster (internal) compared to other clusters (external).
  Suppose there are n POIs $v_1, v_2, ..., v_n$ with coordinates are $x_1, x_2, ..., x_n$, the silhouette value $s(v_i)$ of a POI v_i ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Specifically, $s(v_i)$ close to 1 represents the point is clustered properly and different from other clusters; $s(v_i)$ close to 0 represents the point is not distinctly in one cluster or others; $s(v_i)$ close to -1 represents the point probably assigned to the wrong cluster.
  Meanwhile, the average silhouette value of a cluster (region) measures whether the data within the individual cluster are well grouped, and average silhouette over all the clusters (regions) measures whether all the data are well grouped. In this case, we can calculate the silhouette value within regions, also the average silhouette value over all the regions.
  Note that, for ease of comparison, we normalize the *Silhouette Value* into [0,1], and combine it with *Kernel Density Estimation* together as the metric of the spatial part.

B. Case Study - Functional Region Annotation

In this section, we aim to assign each region with a thematic topic for a better understanding of its real functions. Actually, the region annotation problem is the visualization process, which is listed as a feature distribution of topic modeling in Blei [3]. Listing the most frequent words of a discovered topic to annotate a document in the corpus is a widely used method, but it is not sufficient to summarize the functional regions. We conduct the functional region annotation considering the two parts.

POI Frequency We compute an average POI density vector across the regions in the functional region, where the density ρ_j of the jth POI category in the region r_i is calculated by:

$$\rho_j = \frac{\text{Number of POIs of the } jth \text{ POI category}}{\text{Area of region } r_i} \tag{8}$$

According to the density value of each POI category in the calculated POI density vector, we rank the POI category in a functional region.

Government Open Resource[2] The Los Angeles government puts their city plan on the website for residents or visitors. People may comprehend a few number of functional regions according to the city plan. For example, the region containing the Hollywood City should be an area of movie city.

[2]http://zimas.lacity.org/

IV. EXPERIMENT

A. Experiment Setup

Dataset Our experimental data set is crawled from Twitter[3]containing both semantic information and spatial information, represented as *User Check-in Records* and *POIs*, respectively. In this work, we select both Los Angeles and Manhattan as the study areas, considering both *User check-in Records* and *POIs*.

1 User Check-in Records User check-in records contain the user-ID, check-in time, POI geo-location, POI-ID, and other information, such as check-in city, check-in country, etc. We combine the User Check-in Records and POIs according to the unique POI ID.

Table.I shows some examples of user check-in records.

TABLE I: Examples of User Check-in Records

User ID	POI longitude/latitude	Check-in Time	POI ID
1	40.74359714, -74.00332704	Fri Feb 01 00:08:13 +2016	4bf58dd8d48988d116941735
2	40.78587031, -73.97852922	Sat May 05 02:50:33 +2016	4bf58dd8d48988d147941735
...	...	...	
10	40.80986581, -73.95154698	Wed Jun 06 14:47:53 +2016	4bf58dd8d48988d175941735

2 POIs The POI data contains the POI geo-location, unique POI_ID, POI category words, and other information like POI administrative region. The raw POI include 11 categories, including *[C1] Art & Entertainment*, *[C2] College & University*, *[C3] Event*, *[C4] Food*, *[C5] Nightlife Spots*, *[C6]Outdoors & Recreation*, *[C7] Professional & Other-Places*, *[C8] Residence*, *[C9] Shop & Services*, *[C10] Travel & Transport*.

Table.II shows some examples of POIs.

TABLE II: Examples of POIs

POI ID	POI longitude	POI latitude	POI categorical words
4bf58dd8d48988d116941735	-74.00332704	40.74359714	Bathtub Gin/Bar/Nightlife
4bf58dd8d48988d147941735	-73.97852922	40.78587031	Dinner/Food
...	...	...	...
4bf58dd8d48988d175941735	-73.95154698	40.80986581	Gym/Fitness Center

Parameter Settings All methods are implemented in Java 1.7 and conducted on 12G main memory machine with an Intel 3.4GHz I7-2600 CPU with four cores.

B. Comparative Approaches

K-means Clustering Algorithm K-means clustering algorithm is a method of vector quantization, originally from signal processing. Since it is widely used in the clustering problem, we apply it as one of the baseline methods. We only consider the spatial information, ignore the semantic meaning of each POI. We adopt *Euclidean distance* to measure the performance of each functional region.

LDA Topic Model We only use the semantic information (POI data) as the input data, and the output data is the distribution of region-topic. Each row represents the semantic meaning of regions, and each element represents the probability that the region belonging to each topic.

TF-IDF based Clustering Methods TF-IDF (term frequency-inverse document frequency) [11] is used to measure how important a word is to a document in a corpus. Our target problem is discovering a series of regions with different semantic meanings. In other words, we should measure the importance of a POI in a region. Concretely, for a given region, we calculate the POI importance in the format of a POI importance vector $< v_1, v_2, ..., v_r >$ according to (9):

$$v_i = \frac{n_i}{N} \times \log \frac{R}{\parallel r|i-th \text{ POI category } \in r \parallel} \quad (9)$$

In the TF part, n_i denotes the number of POIs belonging to i-th category and N is the number of POIs located in this region. While in the IDF part, r denotes the numbers of regions containing i-th POI category, R is the number of functional regions.

In this work, we segment the whole city into several small regions according to [18]. Then, we apply (9) to calculate the POI importance vectors, which are important for functional region discovery. Finally, according to [18], we adopt K-means to arrange the regions into 30 functional regions.

DRoF (Discovering Regions of different Functions) Discovering Regions of Functions [18] combines the LDA (Latent Dirichlet Allocation) and DMR (Dirichlet Multinomial Regression) to find the regions with different functions combing user trajectory data and POIs. Similar to the TF-IDF method, this method also arranges the 59 initial regions into 30 functional regions using the K-means clustering algorithm.

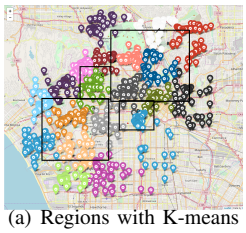
For ease of comparison, we set the number of regions as 30 in all the baseline methods and our proposed method.

C. Results of Functional Region Identification

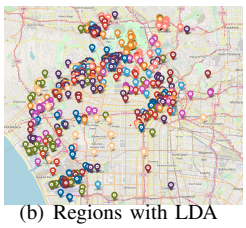
1) Functional Regions Discovery: Fig.2 shows the functional regions of Los Angeles discovered by different methods, and different colors indicate different functions. In different methods, the same color may stand for different functions.

The K-means method only considers the POI geo-location. As shown in Fig. 2(a), the results only reflect the spatial feature, which means the final regions are closely in the spatial space, but the functionality within regions may not semantic meaningful. Basically, the LDA-based method only considers the POI words. Fig. 2(b) shows the regions are diverse in the space. Though the topic of each region is well described, the regions are overlapped in the space. Fig. 2(c) shows the regions with TF-IDF methods. We adopt the existed work [11], using the POI distribution as feature vectors. Fig. 2(d) shows the regions with DRoF method, which is similar to the TF-IDF method. The regions do not have overlapping parts and have different topics after the aggregation step. Fig. 2(e) presents the identified functional

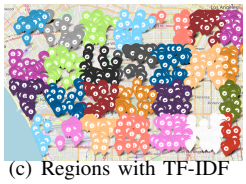
[3]https://developer.twitter.com/



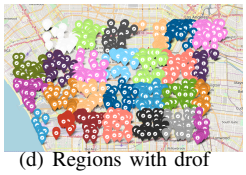
(a) Regions with K-means



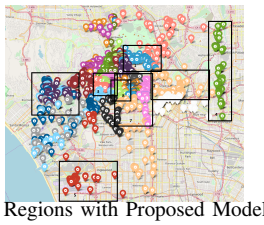
(b) Regions with LDA



(c) Regions with TF-IDF



(d) Regions with drof



(e) Regions with Proposed Model

Fig. 2: Functional Regions with Comparative Models

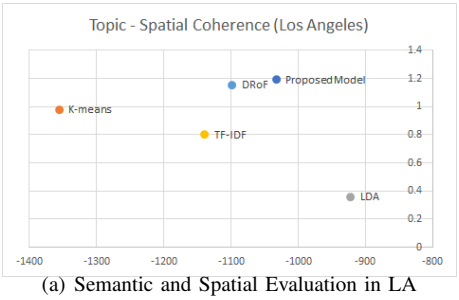
regions using the proposed model combing both semantic words and the spatial information of POIs.

According to the final results, the proposed model can identify the functional regions effectively with considering both the semantic and spatial information. Besides, some miss-identified functional regions can be corrected by incorporating diverse check-in records. For example, region A in Fig. 2(a) is a university region, which is wrongly divided and clustered into the surrounding regions. Region B in Fig. 2(a) is a famous movie city (Hollywood), which should be clustered in an individual region. Region C, D, E are all commercial areas, which are all clustered into a big region according to the government open resource area.

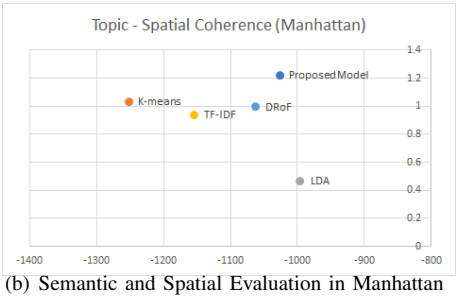
2) Semantic and Spatial Evaluation of Functional Regions: Fig.3 shows our method performing better than all the other methods besides LDA considering the topic coherence, because it only considering the semantic information. The silhouette number of our method close to 0 reflects there are some overlapping regions, which fit the intuition the functional regions are overlapped. Meanwhile, the topic coherence is higher than other baseline methods, representing the thematic meaning of each region are well concluded.

D. Case Study - Functional Region Annotation

Fig.III shows the most frequent words within each regions. Fig.IV shows the average POI frequency, and corresponding internal and external rankings, where the external rank is represented by the depth of the color. Especially, $region1$, $region2$, $region3$, $region7$, $region13$, $region19$ are more



(a) Semantic and Spatial Evaluation in LA



(b) Semantic and Spatial Evaluation in Manhattan

Fig. 3: Semantic and Spatial Evaluation

TABLE III: Most Frequent Words in Each Regions

	Top-5 Categorical Words
Region1	ThemePark,TacaoPlaces, DogRun, KoreanRestaurant, AsianRestaurant
Region2	GayBar,RockClub,NightlifeSpots, Bar, AmericanRestaurant
Region3	Neighborhood,WingJoint,BurgerJoint, MexicanRestaurant,FastFood
Region7	MusicVenue,Neighborhood,CafeStore, GroceryStore,GasStation
Region13	AirportTerminal,BusStation, GeneralTravel,Fastfood,GroceryStore
Region19	CollegeFootballField, CollegeResidence, CollegeAcademicBuilding,ollegeTrack, JuiceBar

TABLE IV: Overall POI Frequency of Functional Regions

	[C1]	[C2]	[C4]	[C5]	[C6]	[C7]	[C8]	[C9]	[C10]
1	0.0308	0	0.028	0.0308	0.0598	0.0218	0.0434	0.0308	0.0451
2	0.0772	0.0348	0.0256	0.1674	0.0176	0.0062	0.0186	0.0132	0
3	0.0308	0	0.0534	0.0837	0.0176	0.0124	0.0062	0.0286	0.0064
4	0.0733	0.0232	0.0367	0.04405	0.0105	0.0342	0.0124	0.0132	0.0451
5	0.0193	0.0232	0.0267	0.0044	0.0563	0.0249	0.0372	0.0418	0.0612
6	0.0270	0	0.0445	0.0132	0.0176	0.0623	0.0186	0.0242	0.0354
7	0.0270	0	0.0267	0.0308	0.0669	0.0311	0.0372	0.0308	0.0419
8	0.1081	0.0465	0.0233	0.0132	0.0352	0.02492	0.0372	0.0242	0.02903
9	0.0501	0.0348	0.0289	0.0088	0.0211	0.0311	0.0434	0.0572	0.0225
10	0.0501	0.0813	0.0267	0.0176	0.0176	0.0373	0.0496	0.0286	0.0451
11	0.0193	0.0581	0.0367	0.0132	0.0387	0.0404	0.0124	0.0286	0.0483
12	0.0154	0.0348	0.0311	0.0132	0.0457	0.0249	0.0434	0.0682	0.0096
13	0.0270	0.0232	0.0345	0.0352	0.0105	0.0342	0.0248	0.0176	0.0838
14	0.0463	0.0116	0.0334	0.0352	0.0316	0.0186	0.0496	0.0506	0.0096
15	0.0270	0.0232	0.0389	0.1101	0.0140	0.0342	0.0186	0.0154	0.0193
16	0.0231	0.0697	0.0178	0.0088	0.0352	0.0342	0.0683	0.0506	0.0483
17	0.0154	0.03489	0.0267	0.0088	0.0387	0.0716	0.0310	0.0220	0.0580
18	0.0424	0.0348	0.0378	0.0308	0.0281	0.0404	0.0186	0.0374	0.0129
19	0.0154	0.1628	0.0356	0.0132	0.0070	0.0467	0.0559	0.0198	0.0387
20	0.0231	0.0116	0.0211	0.0132	0.0563	0.0467	0.0496	0.0638	0.0096
21	0.0308	0.0116	0.0434	0.0132	0.0105	0.0186	0.0621	0.0330	0.0483
22	0.0386	0.1046	0.0467	0.0220	0.0211	0.0155	0.0372	0.0352	0.0032
23	0.0270	0	0.0233	0.0484	0.0281	0.0560	0.0310	0.0198	0.0677
24	0.0193	0.0465	0.0345	0.0176	0.0316	0.0311	0.0559	0.0154	0.0677
25	0.0193	0.0232	0.0489	0.0220	0.0352	0.0311	0.0372	0.0374	0.0032
26	0.0077	0	0.0322	0.1189	0.0422	0.0093	0.0186	0.0264	0.0387
27	0.0308	0.0116	0.0189	0.0044	0.0845	0.0623	0.0248	0.0264	0.0419
28	0.0424	0.0232	0.0345	0.0220	0.0387	0.0373	0.0124	0.0506	0.0096
29	0.0154	0.0232	0.0434	0.0220	0.0387	0.0155	0.0186	0.0572	0.0161
30	0.0193	0.0465	0.0378	0.0132	0.0422	0.0436	0.0248	0.0308	0.0322

mature compared to other regions, since they have more high ranked POI frequency, which are annotated as follows:

Recreation Region[region1] It contains *Theme Park*, *Tacao Places*, *DogRun*, which are correlated with Recreation or Outdoors.

Nightlife Region[region2] This regions contains many places with categorical words *Pub*, *Rock Club* and *neighborhood*. Compared to the government open resource, there are Starbucks, Gastropuds and other nightlife places. Intriguingly, a lot of *Gay Bars* are also grouped into this regions since they sharing similar topic and close in reality.

Entertainment Region[region3] It is a typical entertainment regions containing high probability categorical words, such as *Neighborhood, Wings Joint* and *Korean Restaurant, Mexican Restaurant*.

Movie City Region[region7] This region has the highest POI frequency of *Music Venue*, also *Neighbors, Cafe Store* and *Grocery Store*. Compared to the real map, this area contains the famous movie area Hollywood in LA, which fits the intuition of our proposed method.

Transport Region[region13] POIs in this area are probably *Airport Terminal, Bus Station* and *Airport Port*, Meanwhile, there are some *Fast food restaurant* and *Grocery Store* in this area. In real life, there are some grocery stores and some convenient shops near the airports, which fits humans' common sense.

Educational Region[region19] This kind of regions contains several *College Research Building, College Residence* and *College Teaching Building*, which reflects this region is correlated with education.

V. Conclusion and Future Work

In this paper, we propose a model to automaticaaly discover the functional regions in the modern city using user check-in records and POIs. Specifically, each functional region have different functions with overlapped parts, which is the main difference with the existing methods. Functional regions discovery benefits several applications, such as places recommendation, city plan, and route search. We evaluate our proposed model with the large scale real world dataset, for better understanding the functions of each regions, we conduct a series of case studies. The experiment result shows our proposed method performing better than all the other baseline methods, and the case study illustrates our proposed method is useful and meaningful.

VI. Acknowledgements

This work was partially supported by Shenzhen Fundamental Research Fund (JCYJ20170306141038939, KQJSCX20170728162302784, ZDSYS201707251409055), and Guangdong Introducing Innovative and Entrepreneurial Teams Fund (2017ZT07X152), China.

References

[1] J. Antikainen, "The concept of functional urban area," *Informationen zur Raumentwicklung*, vol. 7, pp. 447–452, 2005.

[2] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] W. Chen, H. Yin, W. Wang, L. Zhao, W. Hua, and X. Zhou, "Exploiting spatio-temporal user behaviors for user linkage," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 517–526.

[5] Y. Ge, C. Liu, H. Xiong, and J. Chen, "A taxi business intelligence system," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 735–738.

[6] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 899–908.

[7] C. Karlsson *et al.*, "Clusters, functional regions and cluster policies," *JIBS and CESIS Electronic Working Paper Series (84)*, pp. 1010–1018, 2007.

[8] J. Miller, "Analysis of fastest and shortest paths in an urban city using live vehicle data from a vehicle-to-infrastructure architecture," *IFAC Proceedings Volumes*, vol. 42, no. 15, pp. 544–548, 2009.

[9] 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*. Association for Computational Linguistics, 2011, pp. 262–272.

[10] 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 *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 384–388.

[11] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, 2003, pp. 133–142.

[12] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[13] R. R. Vatsavai, E. Bright, C. Varun, B. Budhendra, A. Cheriyadat, and J. Grasser, "Machine learning approaches for high-resolution urban land cover classification: a comparative study," in *Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications*. ACM, 2011, p. 11.

[14] M. P. Wand and M. C. Jones, *Kernel smoothing*. Chapman and Hall/CRC, 1994.

[15] W. Wang, H. Yin, L. Chen, Y. Sun, S. Sadiq, and X. Zhou, "Geosage: A geographical sparse additive generative model for spatial item recommendation," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1255–1264.

[16] Y. Wang, Y. Gu, M. Dou, and M. Qiao, "Using spatial semantics and interactions to identify urban functional regions," *ISPRS International Journal of Geo-Information*, vol. 7, no. 4, p. 130, 2018.

[17] H. Yin, B. Cui, L. Chen, Z. Hu, and C. Zhang, "Modeling location-based user rating profiles for personalized recommendation," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 9, no. 3, p. 19, 2015.

[18] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 186–194.