Discovering Regions of Different Functions in a City Using Human Mobility and POIs

Jing Yuan
Microsoft Research Asia
v-jinyua@microsoft.com

Yu Zheng
Microsoft Research Asia
yuzheng@microsoft.com

Xing Xie
Microsoft Research Asia
xing.xie@microsoft.com

ABSTRACT

The development of a city gradually fosters different functional regions, such as educational areas and business districts. In this paper, we propose a framework (titled DRoF) that **D**iscovers **R**egions **o**f different **F**unctions in a city using both human mobility among regions and points of interests (POIs) located in a region. Specifically, we segment a city into disjointed regions according to major roads, such as highways and urban express ways. We infer the functions of each region using a topic-based inference model, which regards a region as a document, a function as a topic, categories of POIs (e.g., restaurants and shopping malls) as metadata (like authors, affiliations, and key words), and human mobility patterns (when people reach/leave a region and where people come from and leave for) as words. As a result, a region is represented by a distribution of functions, and a function is featured by a distribution of mobility patterns. We further identify the intensity of each function in different locations. The results generated by our framework can benefit a variety of applications, including urban planning, location choosing for a business, and social recommendations. We evaluated our method using large-scale and real-world datasets, consisting of two POI datasets of Beijing (in 2010 and 2011) and two 3-month GPS trajectory datasets (representing human mobility) generated by over 12,000 taxicabs in Beijing in 2010 and 2011 respectively. The results justify the advantages of our approach over baseline methods solely using POIs or human mobility.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: data mining, spatial databases and GIS.

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

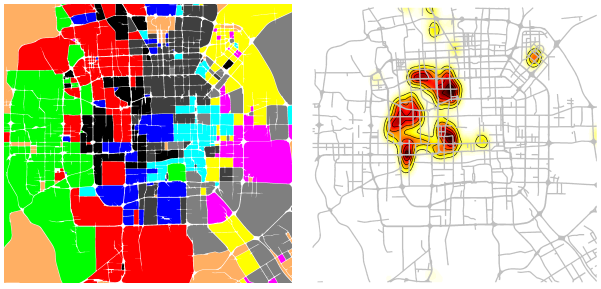
functional regions, urban computing, taxi trajectories, human mobility

1. INTRODUCTION

ACM KDD '12 Beijing, China

The step of urbanization and modern civilization leads to different functional regions in a city, e.g., residential areas, business districts, and educational areas, which support different needs of people's urban lives and serve as a valuable organizing technique for framing detailed knowledge of a metropolitan. These regions may be artificially designed by urban planners, or naturally formulated according to people's actual lifestyle, and would change functions and territories with the development of a city.

In this paper, we aim to discover regions of different functions in urban areas using human mobility and points of interests (POIs). Here, a city is partitioned into individual regions by major roads, like high way and ring roads (refer to Figure 1(a)). Human mobility is represented by people's movement trajectories, which can be cell-tower traces in a cellular network, or trajectories of driving routes, or a sequence of posts (like geo-tweets, geo-tagged photos, or check-ins) in location-based services [21]. A POI is associated with a coordinate and a category like restaurants and shopping malls. Specifically, we fill regions that could have similar functions with the same color in Figure 1(a) and identify the functionality intensity of each function in different locations. For example, Figure 1(b) shows the functionality intensity of developed commercial/entertainment (a kind of function) areas in Beijing, where the darker part suggests a higher intensity. This step is motivated by the following observations. Sometimes, only a part of a region contributes to a function. On the other hand, a function could be formulated across several individual regions (e.g., a shopping street). Finally, each function is titled with some tags in a semi-manual way based on the output of our method.



(a) (b)

Figure 1: The objectives of this paper: a) functional regions; b) intensity of a function

Discovering regions of different functions can enable a variety of valuable applications. First, it can provide people with a quick understanding of a complex city (like New York City, Tokyo, and Paris) and social recommendations. For example, tourists can easily

differentiate some scenic areas from business districts given these functional regions, thereby reducing effort for trip planning. Local people can also expand their knowledge about a city by finding regions that have similar functions (e.g., entertainment areas). It is very common that local people may not well understand each part of a metropolitan even if they have been in the metropolitan for a few years. Second, these functional regions can calibrate the urban planning of a city and contribute to the future planning to some extent. It is not surprising that a city did not evolve as its original planning, given the complexity of urban planning itself and the difficulty in predicting the development of a city. Third, these functional regions would also benefit location choosing for a business and advertisement. For instance, when building a supermarket we need to consider the distance to the residential areas, and the advertisement for a training course could be better put considering the geospatial intensity of the educational function.

To identify the functions of a region, we need to take into account both POIs located in a region and human mobility among these regions due to the following two aspects:
1) *POI data*: On the one hand, POIs feature the function of a region. For example, a region containing a number of universities and schools has a high probability to be an educational area. On the other hand, a region usually contains a variety of POIs, thereby having compound functions instead of a single function. Some regions may serve as both business districts and entertainment areas in a city. In addition, the information from POI data cannot differentiate the quality of different venues and reflect the interactions between functional regions. For instance, restaurants are everywhere in a city, however, they could denote different functions. Some small restaurants were built just for satisfying local residences' daily needs, while a few famous restaurants attracting many people might be regarded as a feature of an entertainment area. As a result, sometimes two regions sharing a similar distribution of POIs could still have different functions.
2) *Human mobility*: The functions of a region have a strong correlation with the traveling behavior of people who visit the region. The knowledge that human mobility contributes to reveal the functions of a region mainly lies in two folds. One is when people arrive at and leave a region. The other is where people come from and leave for. Intuitively, in a workday people usually leave a residential area in the morning and return in the evening. The major time when people visit an entertainment area, however, is the evening of workdays or the whole day of non-workdays. Furthermore, regions of different functions are correlated in the context of human mobility. For instance, people reaching an entertainment area have a high probability from a working area (in a workday) and a residential area (in non-workdays). As a result, two regions are more likely to have similar functions, if people travel to the two regions from similar functional regions or leave for similar ones.

The research reported in this paper is a step towards urban computing which enables smart cities by understanding urban dynamics. The contribution of this paper consists of three points:

- We propose a topic model-based method to identify the functions of individual regions, which are obtained using morphological image segmentation approach. The proposed method regards a region as a document, deems a function as a topic, uses human mobility related to the region as words, and treats POIs located in a region as metadata (like titles, authors, affiliations, and key words). As a result, a region is represented by a distribution of topics (functions), and each topic is a denoted by a distribution of words. This model fits the research intuitively and reduces the data sparseness problem.
- We infer the territory of these functions by clustering regions into groups according to the topic distribution of each region. Regions from the same cluster have similar functions, and different clusters represent different functions. Then, for each cluster of regions, we identify the functionality intensity in the regions (belonging to the cluster), using Kernel Density Estimation which employs human mobility as samples.
- We evaluated our method using large-scale and real-world datasets, consisting of two POI datasets of Beijing (in year 2010 and 2011) and two 3-month GPS trajectory datasets (representing human mobility) generated by over 12,000 taxicabs in Beijing in 2010 and 2011 respectively. The results justify the advantages of our approach over baseline methods solely using POI or human mobility. In addition, these powerful datasets allow us to study not only the functional regions in a city but also the evolving of the city across years.

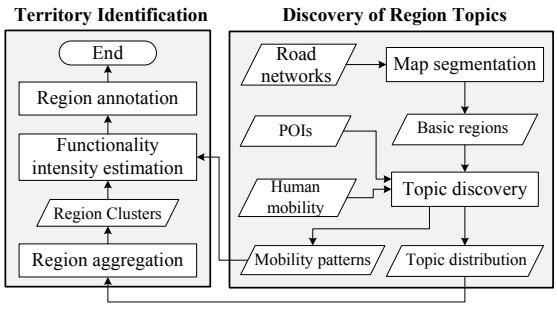


Figure 2: Framework for discovering the functional regions

In accordance with the framework of our work presented in Figure 2, the rest of this paper is organized as follows: Section 2 discovers the distributions of functions for each region, which consists of a map segmentation and an analogy from topic model of documents. Section 3 identifies the aggregated functional regions based on the discovered distribution of functions pertaining to each region and estimates the intensity of each function. Evaluation results are reported in Section 4 and related works are categorized in Section 5. Finally, we briefly conclude the paper in Section 6.

2. DISCOVERY OF REGION TOPICS

In this section we first segment the urban area of a city into region units in terms of major road networks, and then infer the distribution of functions in each region unit using a topic-model-based method.

2.1 Map Segmentation

A road network is usually comprised of some major roads like highways and ring roads, which naturally partition a city into regions. For example, as shown in Figure 4, the red segments denote highways and city express ways in Beijing, and blue segments represent urban arterial roads. The three kinds of roads are associated with a road level 0, 1, and 2 respectively (in a road network database), forming a nature segmentation of the urban area of Beijing. We term each segmented region as a "formal region" in the rest of this paper, following the definition proposed in [2]. Intrinsically, a formal region is a basic unit carrying social-economic functions due to the following reasons. First, people live in formal regions and POIs fall in regions. Second, formal regions as the origin and destination of a trip are the root cause of human mobility. In short, people travel among formal regions.

In our method, we choose the raster-based model to represent the road network and utilize morphological image processing techniques to address the task of map segmentation. Typically, in a

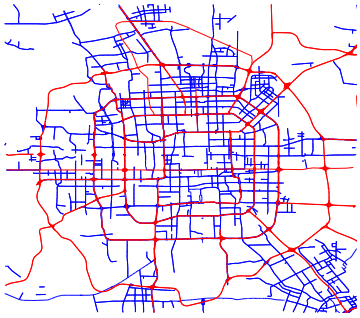


Figure 4: Beijing road network. red: level-0/1; blue: level-2

Geographical Information System (GIS), there are two models to represent spatial data: *vector*-based model and *raster*-based model. Vector-based model uses geometric primitives such as points, lines and polygons to represent spatial objects referenced by Cartesian coordinates, while raster-based model quantizes an area into small discrete grid-cells. Both of the two models have advantages and disadvantages depending on the specific applications. For instance, vector-based method is more powerful for precisely finding shortest-paths, whereas it requires intensive computation when performing topological analysis, such as the problem of map simplification[5], which is proved to be NP-complete [5]. On the other hand, raster-based model is more computational efficient and succinct for territorial analysis, but the accuracy is limited by the number of cells used for discretizing the road networks.

Specifically, a raster-based map is a binary image (e.g., 0 stands for road segments and 1 stands for blank space). In order to remove the unnecessary details, such as the lanes of a road and the overpasses (see Figure 3(a)), we first perform a *dilation* operation to thicken the roads. As a result,we can fill the small holes and smooth out the unnecessary details, as shown in Figure 3(b). Second, we obtain the skeleton of the road networks by performing a *thinning* operation based on the algorithm proposed in [8], as depicted in Figure 3(c). This operation recovers the size of a region which was reduced by the dilation operation, while keeping the connectivity between regions. The last step is to perform a connected component labeling (CCL) that finds individual regions by clustering "1"-labeled grids, using the method proposed in [14]. Figure 3(d) presents the results.

2.2 Topic Discovery

2.2.1 Preliminary

DEFINITION 1 (TRANSITION). A *transition* Tr is a quaternion containing the following four items: origin region ($Tr.r_O$), leaving time ($Tr.t_L$), destination region ($Tr.r_D$) and arrival time ($Tr.t_A$). Here, $Tr.r_O$ and $Tr.r_D$ are spatial features while the others are temporal features. □

DEFINITION 2 (MOBILITY PATTERN). A *mobility pattern* M is a triple extracted from a transition. Given a transiton $Tr = (Tr.r_O, Tr.r_D, Tr.t_L, Tr.t_A)$, we obtain two mobility patterns: the *leaving mobility pattern* $M_L = (Tr.r_O, Tr.r_D, Tr.t_L)$, and the *arriving mobility pattern* $M_A = (Tr.r_O, Tr.r_D, Tr.t_A)$. □

DEFINITION 3 (TRANSITION CUBOIDS). A *transition cuboid* C is an $R \times R \times T$ cuboid, where R is the number of regions and T is the number of time bins. Since there exist two types of mobility patterns, we define two types of transition cuboids: *leaving cuboid* C_L and *arriving cuboid* C_A. The cell with index (i, j, k) of the leaving cuboid records the number of mobility patterns that leave r_i for r_j at time t_k, i.e.,

$$C_L(i, j, k) = \|\{M_L = (x, y, z) | x = r_i, y = r_j, z = t_k\}\|.$$

Similarly,

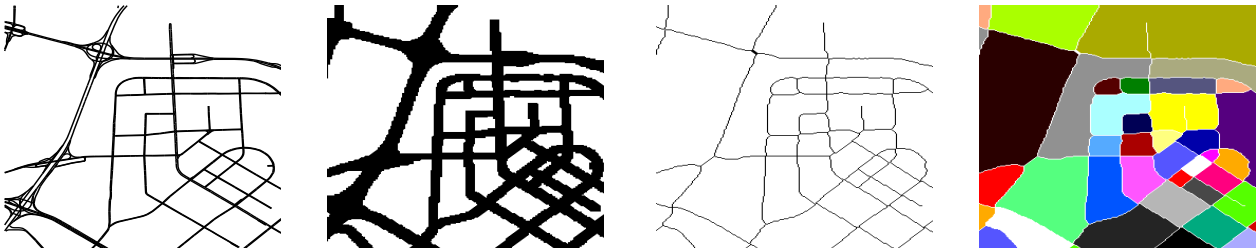
$$C_A(i, j, k) = \|\{M_A = (x, y, z) | x = r_i, y = r_j, z = t_k\}\|. \quad \square$$

We project each trajectory representing human mobility on the segmented region units, turning a trajectory into a transition. Then, we discretize time of day into time bins in each of which we deposit the transitions and formulate mobility patterns. Here, we do not differentiate different weekdays but differ the time bins in weekdays from those in weekends. For example, setting 2 hours as a bin, we will have 24 bins (12 for weekdays and 12 for weekends) in total. Later, two transition cuboids are built using the mobility patterns.

Concepts of Topic Models. Probabilistic topic models have been successfully used for extracting the hidden thematic structure in large archives of documents[3]. In this model, each *document* of a *corpus* exhibits multiple *topics* and each *word* of a document supports a certain topic. Given all the words of each document in a corpus as observations, a topic model is trained to infer the hidden thematic structure behind the observations. Latent Dirichlet Allocation (LDA) is a generative model that includes hidden variables. The intuition behind this model is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [4]. Let α and η be the prior parameters for the Dirichlet document-topic distribution and topic-word distribution respectively. Assume there are K topics and β is a $K \times V$ matrix where V is the number of words in the vocabulary (all the words in the corpus D). Each β_k is a distribution over the vocabulary. The topic proportions for the dth document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in the dth document. The topic assignments for the dth document are z_d, where $z_{d,n}$ is the topic assignment for the nth word in the dth document. Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary.

Using the above notations, as presented in Figure 5, the generative process can be described as follows:

1. For each topic k, draw $\beta_k \sim Dir(\eta)$.



(a) Original (b) Step 1: dilation (c) Step 2: thinning (d) Step 3: CCL

Figure 3: Map Segmentation

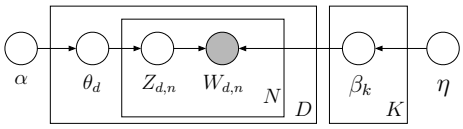


Figure 5: Graphic model of LDA

2. Given the dth document d in corpus D, draw $\theta_d \sim Dir(\alpha)$.
3. For the nth word in the dth document $w_{d,n}$,
   (a) draw $z_{d,n} \sim Mult(\theta_d)$;
   (b) draw $w_{d,n} \sim Mult(\beta_{z_{d,n}})$.

Here, $Dir(\cdot)$ is the Dirichlet distribution and $Mult(\cdot)$ is the multinomial distribution. The estimation of LDA can be implemented using the EM algorithm and the most commonly used inference method of LDA is Gibbs sampling. See [4] for a variational inference method and detailed discussions of LDA.

2.2.2 *Topic Modeling*

As shown in Table 1, we draw an analogy between discovering functions of a region and the topic discovery of a document. Specifically, we regard a formal region as a document and a function as a topic. In other words, a region having multiple functions is just like a document containing a variety of topics. Meanwhile, we deem the mobility patterns associated with a region as words and POIs as metadata of a document, as a functional region is characterized by its agglomeration of activities, its intraregional transport infrastructure, mobility of people, and inputs within its interaction borders [7].

Table 1: Analogy from region-functions to document-topics

transition cuboids	⟶ vocabulary
formal regions	⟶ documents
function of a region	⟶ topic of a document
mobility patterns	⟶ words
POI feature vector	⟶ metadata of a document

Figure 6 further details the analogy using an example. In our method, given the mobility dataset, we build the arriving and leaving cuboids respectively according to Definition 3. For a specific region r_i, the mobility patterns associated with r_i are counted by $C_A(1{:}R, i, 1{:}T)$ and $C_L(i, 1{:}R, 1{:}T)$, which are two "slices" extracted from arriving cuboid and leaving cuboid (termed as arriving matrix and leaving matrix respectively). The right part of Figure 6 shows a "document" we compose for region r_1, where a cell (in the matrices) represents a specific mobility pattern and the numbers in the cell denote the occurrences of the pattern. For example, in the right most column of the arriving matrix, the cell containing "5" means on average the mobility that went to r_1 from r_j in time bin t_k occurred 5 times per day. A POI is recorded with a tuple (in a POI database) consisting of a POI category (as listed in Table 2), name and a geo-position (latitude, longitude). For each formal region r, the number of POIs in each POI category can be counted. The *frequency density* v_i of the ith POI category in r is calculated by:

$$v_i = \frac{\text{Number of the POIs of the } i\text{th POI category}}{\text{Area of region } r \text{ (measured by grid-cells)}},$$

and the POI *feature vector* of r is denoted by $x_r = (v_1, v_2, \ldots, v_F, 1)$ where F is the number of POI categories and the last "1" is a default feature to account for the mean value of each topic, as explained in

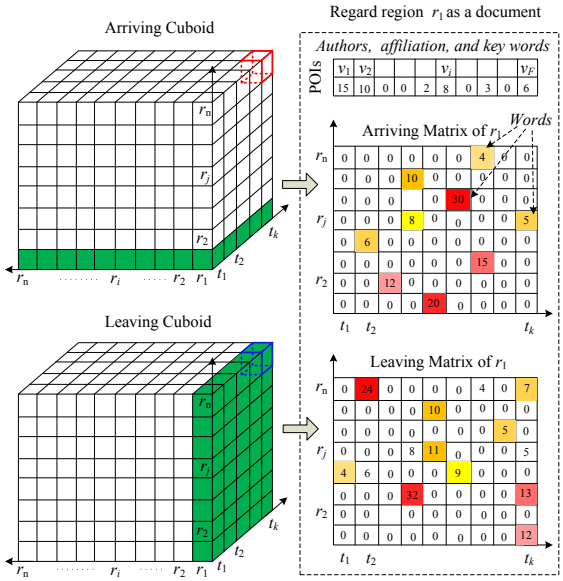


Figure 6: Analogy between mobility patterns and words based on transition cuboids

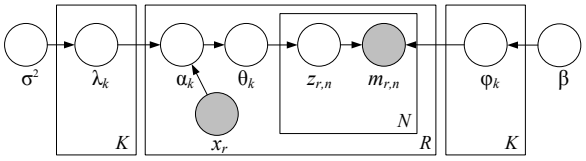


Figure 7: DMR-based topic model

[10]. The POI feature vector is regarded as the metadata of each region, which is an analogue of the observed features such as author/email/institution of a document.

Using the basic LDA model, region topics can be discovered using the mobility patterns. However, as stated in Section 1, the region topics are products of both the POIs and mobility patterns. In order to combine the information from both of them, we utilize a more advanced topic model based on LDA and Dirichlet Multinomial Regression (DMR) [10].

The DMR-based topic model (for simplicity, DMR in the rest of the paper) takes into account the influence of the observable metadata in a document by using a flexible framework, which supports arbitrary features [10]. Compared to other models designed for specific data such as Author-Topic model and Topic-Over-Time model (a member in the supervised-LDA family of topic models), DMR achieves similar or improved performance while is more computational efficient and succinct in implementation [10].

As presented in Figure 7, the generative process of the DMR model is:

1. For each region topic k,
   (a) draw $\lambda_k \sim \mathcal{N}(0, \sigma^2 I)$;
   (b) draw $\beta_k \sim Dir(\eta)$.
2. Given the rth region,
   (a) for each region topic k, let $\alpha_{r,k} = \exp(x_r^T \lambda_k)$;
   (b) draw $\theta_r \sim Dir(\alpha_r)$;
   (c) for the nth mobility pattern in the rth region $m_{r,n}$,
      i. draw $z_{r,n} \sim Mult(\theta_r)$;
      ii. draw $m_{r,n} \sim Mult(\beta_{z_{r,n}})$.

Here, $\mathcal{N}$ is the Gaussian distribution with σ as a hyper parameter, and λ_k is a vector with the same length as the POI feature vector.

Table 2: POI category taxonomy

code	POI category	code	POI category
1	car service	16	banking and insurance service
2	car sales	17	corporate business
3	car repair	18	street furniture
4	motorcycle service	19	entrance/bridge
5	Café/Tea Bar	20	public utilities
6	sports/stationery shop	21	Chinese restaurant
7	living service	22	foreign restaurant
8	sports	23	fastfood restaurant
9	hospital	24	shopping mall
10	hotel	25	convenience store
11	scenic spot	26	electronic products store
12	residence	27	supermarket
13	governmental agencies and public organizations	28	furniture building materials market
14	science and education	29	pub/bar
15	transportation facilities	30	theaters

The nth observed mobility pattern of region r is denoted as $m_{r,n}$. Other notations are similar to the previous LDA model. This model can also be estimated using EM and inferred using Gibbs sampling. Different from the basic LDA model, here, the Dirichlet prior α is now specified to individual regions (α_r) based on the observed POI features of each region. Therefore, for different combination of POI category distributions, the resulting α values are distinct. Thus the thematic region topic distributions extracted from the data are induced by both the POI features and mobility patterns. As a result, by applying DMR, given the mobility patterns and POI features, we obtain the topic assignment for each region and the mobility pattern distribution of each topic.

3. TERRITORY IDENTIFICATION

3.1 Region Aggregation

This step aggregates similar formal regions in terms of region topic distributions by performing a clustering algorithm. Regions from the same cluster have similar functions, and different clusters represent different functions. For region r, after parameter estimations based on the DMR model, the topic distribution is a K dimensional vector $\theta_r = (\theta_{r,1}, \theta_{r,2}, \ldots, \theta_{r,K})$, where $\theta_{r,k}$ is the proportion of topic k for region r. We perform the k-means clustering method on the K-dimensional points $\theta_r,\ r \in 1, 2, \ldots, R$. The number of clusters can be predefined according to the needs of an application or determined using the average *silhouette* value as the criterion [13]. The silhouette value of a point i in the dataset, denoted by $s(i)$ is in the range of $[-1, 1]$, where $s(i)$ close to 1 means that the point is appropriately clustered and very distant from its neighboring clusters; $s(i)$ close to 0 indicates that the point is not distinctly in one cluster or another; $s(i)$ close to -1 means the point is probably assigned to the wrong cluster. The average silhouette value of a cluster measures how tightly the data in this cluster are grouped, and the average silhouette of the entire dataset reflects how appropriately all the data has been clustered. In practice, we perform cross validation on the dataset for different k multiple times and choose an appropriate k with the maximum overall silhouette value. Consequently, we aggregate the formal regions into k clusters, each of which is termed as a *functional region*.

3.2 Functionality Intensity Estimation

On the one hand, the functionality of a functional region is generally not uniformly distributed within the entire region. On the other hand, sometimes, the core functional area may span multiple formal regions and may have an irregular shape, e.g., a hot shopping street crossing several formal regions. In order to reveal the degree of functionality and glean the essential territory of a functional region, we estimate the *functionality intensity* for each aggregated functional region (a cluster of formal regions).

Intuitively, the number of visits implicitly reflects the popularity of a certain functional region. In other words, people's mobility patterns imply the functionality intensity. As a result, we feed the origin and destination of each mobility (represented by latitude and longitude) into a Kernel Density Estimation (KDE) model to infer the functionality intensity in a functional region. Note that the real place that an individual visited may not be the destination that we can obtain from a mobility dataset. For example, the drop-off points of taxi trajectories may not be people's final destinations like a shopping mall. However, the pick-up/drop-off points should not be too far from the really-visited locations according to commonsense knowledge. The farther distance a location to the drop-off point, the lower probability that people would visit the location.

Given n points $x_1, x_2, \ldots, x_n$ located in a 2D spatial space, KDE estimates the intensity at location s by a kernel density estimator, defined as:

$$\lambda(s) = \sum_{i=1}^{n} \frac{1}{nr^2} K(\frac{d_{i,s}}{r}), \tag{1}$$

where $d_{i,s}$ is the distance from x_i to s, r is the bandwidth and $K(\cdot)$ is the kernel function whose value decays with the increasing of $d_{i,r}$, such as the Gaussian function, Quartic function, Conic and negative exponential. The choice of the bandwidth usually determines the smoothness of the estimated density – a large r achieves smoother estimation while a small r reveals more detailed peaks and valleys. In our case, we choose the Gaussian function as the kernel function, i.e.,

$$K(\frac{d_{i,s}}{r}) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{d_{i,s}^2}{2r^2}). \tag{2}$$

and the bandwidth r is determined according to MISE criterion [16].

3.3 Region Annotation

In this step, given the results we obtained, we try to annotate each cluster of regions with some semantic terms, which could contribute to the understanding of its real functions. Note that region annotation is a very challenging problem in both traditional urban planning and document processing. Essentially, this annotation problem is the visualization problem of topic model, which is listed as a future direction of topic modeling in the recent survey paper by Blei [3]. A compromised method so far is to use the most frequent words in a discovered topic to annotate a document. But in our case, listing the frequent mobility patterns (analogue to words) is far from enough to name a functional region.

In our method, we annotate a functional region by considering the following 4 aspects: 1) The POI configuration in a functional region. We compute an average POI feature vector across the regions in functional region. According to the average frequency density in the calculated POI feature vector, we rank each POI category in a functional region (termed as internal ranking) and rank all the functional regions for each POI category (termed as external ranking). We will give an example in the experiment as shown in Table 4. 2) The most frequent mobility patterns of each functional region. 3) The functionality intensity. We study the representative POIs located in each functionality kernel, e.g., a function region could be an educational area if its kernel is full of universities and schools. 4) The human-labeled regions. People may know the functions of

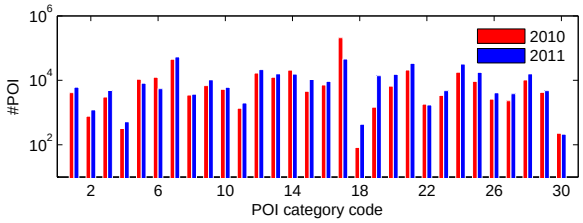


Figure 8: POI counts for each category in 2010 and 2011

a few well-known regions, e.g., the region contains the Forbidden City is an area of historic interests. After clustering, the human labeled regions will help us understand other regions in a cluster. Refer to the experiments for the detailed results and analysis.

4. EVALUATION

4.1 Settings

We use the following datasets for the evaluation:
a) *POI Data*: Two Beijing POI datasets in year 2010 and 2011. The count of each POI category (see Table 2) is presented in Figure 8.
b) *Mobility Data*: Two GPS trajectory datasets generated by Beijing taxis in 2010 and 2011, with the statistics shown in Table 3. We only choose the occupied trips (identified by the information of a taxi meter) from the data. Actually, there are over 30 cities in this world having over 10,000 taxicabs, e.g., Beijing has over 67,000 taxis. The taxi trips represent a significant portion of people's urban mobility. According to the report of Beijing Transportation Bureau, the taxi trips occupy over 12 percent of traffic flows on road surfaces[19]. Of course, other mobility datasets such as mobile phone traces can be used in our framework.
c) *Road Networks*: We have the road networks of Beijing in 2010 and 2011, with statistics shown in Table 3.

Table 3: Statistics of taxi trajectories and road networks

	year	2010	2011
Trajectories	#taxis	12,726	13,597
	#occupied trips	21,678,203	8,202,012
	#effective days	112	92
	average trip distance(km)	7.22	7.47
	average trip duration(min)	15.98	16.1
	average sampling interval(sec)	74.46	70.45
Roads	#road segments	150,357	162,246
	percentage of major roads	18.9%	17.1%
	#segmented formal regions	565	554
	size of "vocabulary" (non-0 items)	3,318,331	3,244,901

We implement our method on a 64-bit server with a Quad-Core 2.67G CPU and 16GB RAM. We train our model with 10 topics for 1000 iterations, optimize the parameters every 50 iterations. For k-means clustering, we incorporate the average silhouette value to determine the k and use the average results based on a 5-fold cross-validation. The efficiency (on average) is presented in Table 4.

Table 4: Overall efficiency of DRoF

operation	time(min)
map segmentation	0.325
building transition cuboids	40.1
estimate topic model(1000 iterations)	1353
region aggregation	0.124
total	1394

We compare our method with two baselines: a) the *TF-IDF-based Clustering*, which solely uses the POI data. This method employs the *term frequency-inverse document frequency* (tf-idf) to measure the importance of a POI in a formal region. Specifically, for a given formal region, we formulate a POI vector, $\langle v_1, v_2, \ldots, v_F \rangle$ where v_i is the tf-idf value of the i-th POI category, given by:

$$v_i = \frac{n_i}{N} \times \log \frac{R}{\|\{r | \text{the } i\text{-th POI category} \in r\}\|},$$

where n_i is the number of POIs belonging to the i-th category and N is the number of POIs located in this region. The idf term is calculated by computing the quotient of the number of regions divided by the number of regions which have the i-th POI category, and taking the logarithm of that quotient. Later, a k-means clustering is used to cluster the formal regions into k functional regions. b) *LDA-based Topic Model*, which uses only the mobility data. Similar to our analogy from regions to documents, this method feeds the mobility patterns (the analogue to words) into an LDA model. Later, we perform the k-means clustering, similar to the method we used when grouping all the formal regions based on their topic distributions learned from LDA. The parameters such as number of iterations, number of topics are set in accordance with our DRoF. As the number of POI categories usually has the same scale with that of the topics, applying the LDA model solely to POIs (as words) will not reduce the dimension of words.

We carry out the following three studies to evaluate the effectiveness of our framework though it is very difficult. 1) We invite some local people (who have been in Beijing for over 6 years) and request them to label two representative regions for each kind of function. We check whether the regions having the same labels are assigned into the same functional region and whether the regions with different labels are improperly clustered into one functional region. 2) We find the evolving of Beijing by comparing the results of 2010 and 2011, and identify whether the differences make sense. 3) We match our results against the land use planning of Beijing.

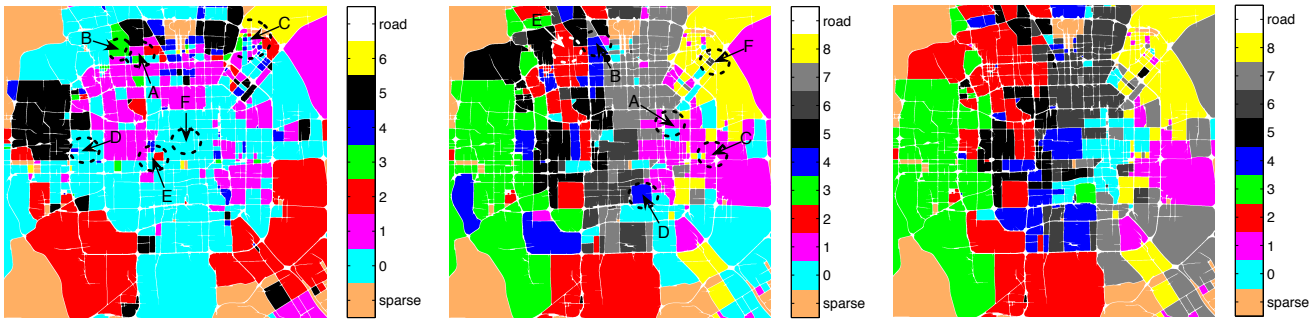
4.2 Results

4.2.1 Discovered Functional Regions

Figure 9 shows the aggregated functional regions discovered by different methods, where different colors indicate different functions. Note that in different figures, the same color may stand for different functions. As a result, tf-idf-based method forms 7 clusters (c_0–c_6) while LDA-based method and our method (DRoF) form 9 functional regions.

The TF-IDF method performs the worst among the three approaches. For example, as shown in Figure 9(a), region B is an university, which should be clustered with region A (another university) and region D (a high school). Meanwhile, region F (the Forbidden City) is not distinguished from other commercial areas like region E (Xidan). Another example is the Wangjing area (C), which is an emerging residential area with some companies and many living services, like apartments, shopping malls and restaurants. Unfortunately, the TF-IDF method improperly divides this area into many small functional regions as this method only considers POI distributions.

Basically, the LDA-based method and DRoF have a similar output of functional regions. However, there still exist several exemplary regions where DRoF outperforms LDA obviously. For example, region F in Figure 9(b) is a developing commercial/entertainment area in Wangjing area. But LDA aggregates it with the Forbidden City (Region F in Figure 9(a)), which is a region of historical interests; Area B (China Agricultural University) and Area E (Tsinghua University) are typical science and education areas where LDA fails to correctly cluster them together; Area A around Sanlitun is a well-known diplomatic district of Beijing, which is mixed



(a) TF-IDF (b) LDA (c) DRoF

Figure 9: Functional regions discovered by different methods

Table 5: Overall POI feature vector and ranking of functional regions by DRoF. FD: frequency density, IR: internal ranking

	c0		c1		c2		c3		c4		c5		c6		c7		c8	
POI	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR
CarServ	0.046	25	0.016	23	0.052	26	0.044	18	0.060	17	0.028	25	0.056	24	0.091	13	0.053	21
CarSale	0.009	28	0.005	27	0.061	24	0.006	27	0.009	27	0.005	28	0.021	27	0.015	26	0.006	27
CarRepa	0.021	26	0.011	24	0.062	23	0.042	19	0.051	20	0.023	27	0.062	23	0.057	18	0.039	25
MotServ	0.002	30	0.003	28	0.004	28	0.001	28	0.002	29	0.004	29	0.001	29	0.001	29	0.003	28
Caf/Tea	0.226	14	0.121	9	0.226	12	0.066	15	0.113	13	0.252	6	0.237	13	0.052	19	0.153	10
StaStor	0.135	17	0.037	20	0.127	17	0.037	20	0.058	18	0.080	19	0.100	19	0.073	15	0.072	17
LivServ	1.289	1	0.581	2	1.322	2	0.399	1	0.698	1	0.780	2	1.345	2	0.430	2	0.886	2
Sports	0.054	23	0.035	21	0.092	21	0.030	22	0.041	22	0.033	23	0.080	20	0.035	20	0.093	16
Hospital	0.244	13	0.088	13	0.222	13	0.069	14	0.137	12	0.144	15	0.246	12	0.070	16	0.194	8
Hotel	0.202	15	0.063	16	0.115	18	0.058	16	0.071	16	0.086	18	0.211	15	0.059	17	0.049	22
SceSpo	0.048	24	0.007	26	0.032	27	0.012	25	0.016	25	0.029	24	0.044	25	0.012	27	0.031	26
Residen	0.795	3	0.230	5	0.638	6	0.203	5	0.323	5	0.398	5	0.797	4	0.221	4	0.440	3
Gov/Pub	0.442	7	0.103	11	0.276	11	0.094	10	0.188	9	0.169	12	0.375	7	0.177	6	0.150	11
Sci/Edu	0.315	11	0.139	7	1.084	3	0.109	9	0.323	6	0.251	8	0.530	6	0.124	9	0.266	6
TrasFac	0.459	6	0.115	10	0.397	7	0.091	11	0.150	11	0.191	11	0.364	8	0.113	10	0.257	7
Bank/Fina	0.376	9	0.128	8	0.383	8	0.078	13	0.107	14	0.197	10	0.320	10	0.083	14	0.135	12
CopBusi	1.128	2	0.593	1	1.947	1	0.334	2	0.348	4	0.548	4	1.738	1	0.475	1	0.977	1
StrFur	0.002	29	0.000	30	0.001	30	0.001	30	0.000	30	0.001	30	0.000	30	0.001	30	0.000	30
Entr/Bri	0.296	12	0.065	14	0.210	14	0.081	12	0.160	10	0.160	14	0.228	14	0.133	7	0.097	15
PubUti	0.405	8	0.101	12	0.285	9	0.112	8	0.238	7	0.209	9	0.314	11	0.132	8	0.132	13
ChiRes	0.692	5	0.252	4	0.926	5	0.294	3	0.399	3	0.813	1	0.829	3	0.235	3	0.370	4
ForRes	0.098	18	0.050	17	0.054	25	0.010	26	0.009	26	0.163	13	0.063	21	0.018	25	0.101	14
FasRes	0.095	19	0.046	18	0.141	16	0.034	21	0.050	21	0.126	16	0.132	17	0.026	22	0.057	20
ShopMal	0.724	4	0.268	3	0.929	4	0.242	4	0.476	2	0.559	3	0.734	5	0.203	5	0.306	5
ConvStor	0.370	10	0.157	6	0.281	10	0.128	7	0.234	8	0.251	7	0.362	9	0.108	11	0.160	9
E-Stor	0.056	21	0.017	22	0.107	20	0.029	23	0.037	23	0.037	22	0.063	22	0.018	24	0.040	23
SupMar	0.055	22	0.008	25	0.065	22	0.020	24	0.025	24	0.042	21	0.040	26	0.021	23	0.040	24
FurBuil	0.086	20	0.065	15	0.151	15	0.192	6	0.093	15	0.088	17	0.142	16	0.099	12	0.064	19
Pub/Bar	0.179	16	0.043	19	0.114	19	0.044	17	0.053	19	0.060	20	0.120	18	0.031	21	0.071	18
Theater	0.011	27	0.001	29	0.002	29	0.001	29	0.006	28	0.025	26	0.007	28	0.002	28	0.002	29

with a developing commercial area C; Region D is a park where LDA fails to put it with other park/mountain areas (the green regions shown in Figure 9(b)). The LDA method only using human mobility overlooks the feature of POIs, thereby drops behind the DRoF (shown in Figure 9(c)).

4.2.2 *Region Annotation*

Table 5 shows the average POI feature vector of each region cluster (c_0–c_8, remember that DRoF generated 9 clusters) and the corresponding internal and external rankings, where the external rank is represented by the depth of the color (1 darkest, 4 lightest). Clearly, region cluster c_0, c_2, c_5, c_6, c_8 are more mature and more developed areas as compared to other clusters, since they have more high ranked POI categories, which are annotated as follows:

Diplomatic and embassy areas[c_0]. The most characteristic POI category in this functional region is the governmental agencies and public organizations, with a significant higher frequency density than other functional regions. Actually, most of the embassies are located in these areas, which are well configured for the diplomatic function, e.g., they have the highest external rank of Pub/Bars and transportation facilities, and the second highest rank of residential buildings, hospitals, and hotels, among all the clusters.

Education and science areas[c_2]. This region cluster contains the maximum number of science and education POIs (e.g., Tsinghua university and Beijing university). In addition, the biggest electronic market in China "ZhongguanCun", known as the Silicon Valley in China, is located in this functional region.

Developed residential areas[c_6]. This region cluster is clearly a mature residential area with the most residential building, living services, hospitals, hotels. In this kind of areas, an adequate number of services supports the people's living, such as the restaurants, shopping malls, banking services, schools, sports centers.

Emerging residential areas[c_8]. This area is annotated as the emerging residential area since it has a balanced POI configuration, such as living services, residential buildings, sports centers, hospitals and some companies etc.

Figure 10 compares the arriving/leaving transitions of (region clusters) c_6 and c_8 with that of other clusters respectively, where the x-axes are time of day (by hour) and y-axes are the functional regions that come from (left subfigures) and leave for (right subfigures). Both c_6 and c_8 follow the trend that most leaving transitions in the morning (8-9am, go to work) and most arriving transitions in the early evening (5-6pm, go back home), which is a typical pattern for the residential area. However, in terms of the absolute quantity, c_8 is much lower than c_6, which shows that there are more people living in c_6.

Developed commercial/entertainment areas[c_5]. They are typical entertainment areas with the highest external rank of theaters, foreign restaurants and café/tea bars. Moreover, there are a great many shopping malls, Chinese restaurant and convenience stores. Figure 12 shows the difference of arriving transitions of c_5 (developed commercial areas) on weekdays and weekends. It's clear that the people reach this kind of areas from the residential regions(c_6) much earlier on weekends (about 9am-11am) than on weekdays (7pm-9pm).

With regard to the other region clusters, since the frequency densities of POIs are much lower than the above functional regions, we identify their semantic functions with more consideration on the functionality intensity and frequent mobility patterns derived for each functional region in addition to the POI configurations.

Developing commercial/business/entertainment areas [c_1]. The POI configuration of this cluster is similar to cluster c_5 and c_7, but in terms of the absolute quantity, c_1 is less than c_5 while more than c_7. A certain number of shopping malls, restaurants and banking services feature this cluster as a developing commercial/ business/ entertainment functional region (either of them is possible). In the meantime, the functionality intensity provides another corroboration for this annotation. As depicted in Figure 13(a), the core of this functional regions is the new CBD of Beijing.

Regions under construction[c_7]. As analyzed above for region c_1, this region will potentially become regions 1 or 8 since the POI configuration produces a rudiment of the commercial/residential area

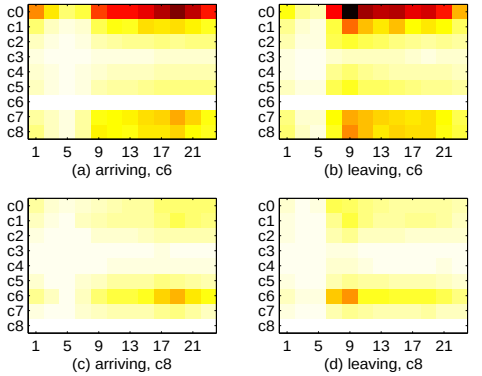


Figure 10: Transitions of c_6 and c_8

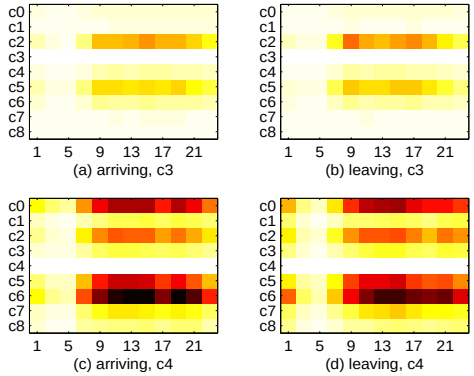


Figure 11: Transitions of c_3 and c_4

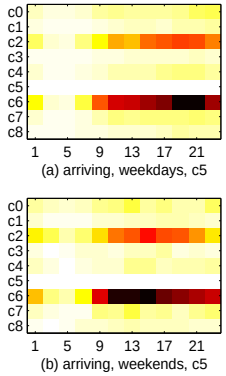
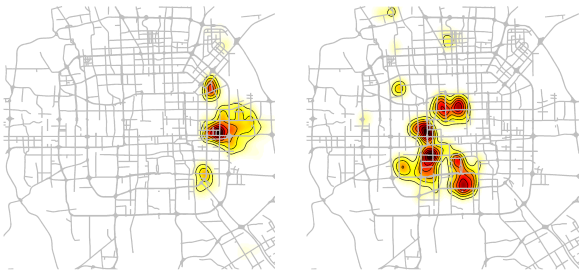


Figure 12: Transitions of c_5



(a) functional region c_1 (b) functional region c_4

Figure 13: Functionality intensity of functional regions

with related supporting services. Figure 14 validates the degree of development with respect to c_5, c_1 and c_7.

Areas of historic interests[c_4]. If we only consider the POI configuration, the characteristic of this cluster does not reveal obviously, which contains some public utilities, entrance/bridges, government organizations, science and education spots. However, by considering the functionality intensity estimated by the mobility patterns, we are easy to find that they are places of historic interests in Beijing. As shown in Figure 13(b), the famous historical places like Forbidden City and Temple of Heaven are located in these areas.

Nature and parks[c_3]. These areas have the fewest POIs in most of the POI categories. Actually, a lot of forests and mountains cover this cluster, e.g., the Xishan Forest Park, Century Forest Park, Baiwang Moutain etc. Figure 11 shows that people come to this functional region following the similar temporal patterns with c_4 (the historical areas), but the diversity and quantity are reasonably weaker than c_4, since many POIs in c_4 are very famous scenic spots.

4.2.3 *Results in Different Years*

We apply our method to the data (road network, POI and mobility) of 2010 and 2011 respectively. The discovered functional regions in 2010 are similar with 2011, while is slightly different in some regions. Figure 15 shows the detailed comparison around the east areas of the Forbidden City. For example, region A (Qianmen Street) becomes a developed commercial area from a nature/park area. The reason is that this region was developing in 2010 after a major repair since 2009. Similar to region A, region B (close to the new CBD of Beijing) becomes a developing commercial area from an under construction area. Intriguingly, region C becomes an under construction area in 2011 from an developing commercial area! We later found the truth that the tallest building of Beijing will rise up in this region in 2015 leading to a relocation work in 2011.

4.2.4 *Calibration for Urban Planning*

The discovered functional regions provide calibration and reference for urban planning. For example, Figure 16 presents the comparison between the governmental land use planning (2002-2010) and the results of DRoF in 2011. This area forms an emerging residential area as planned by the government, while some small regions become developing commercial areas, such as A, B and C after 2 years' development.

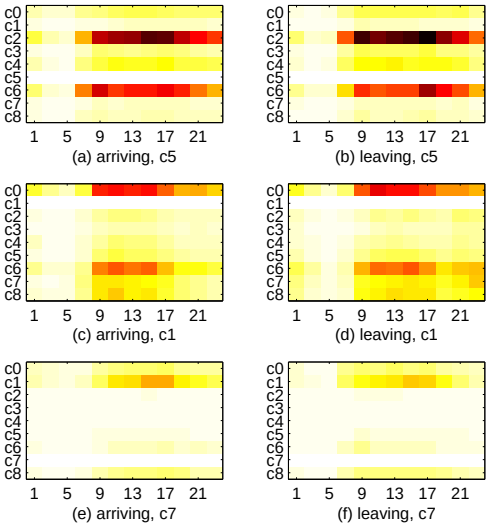


Figure 14: Transitions of c_1, c_5 and c_7

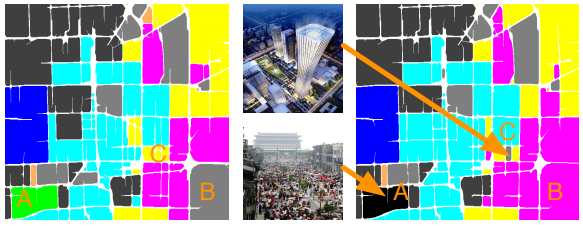
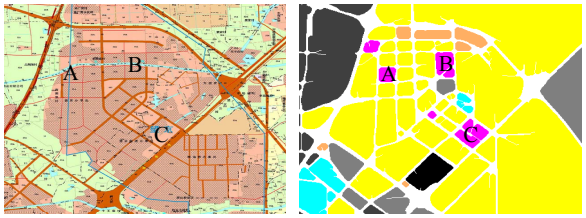


Figure 15: The east area of Forbidden City in 2010 and 2011

5. RELATED WORK

Urban computing with taxicabs: Urban computing is emerging as a concept where every sensor, device, person, vehicle, building, and street in urban areas can be used as a component to probe city dynamics and further enable a city-wide computing for serving people and their cities. The increasing availability of GPS-embedded taxicabs provides us with an unprecedented wealth to understand human mobility in a city, thereby enabling a variety of novel urban computing research recently. For example, Ge et al. [6] and



(a) (b)

Figure 16: (a) governmental land use planning (2002-2010) (b) discovered functional regions in 2011

Yuan et al. [19] respectively study the strategies for improving taxi drivers' income by analyzing the pick-up and drop-off behavior of taxicabs in different locations. [18] aims to find the practically fastest driving route to a destination according to a large number of taxi trajectories, and Zheng et al. [20] glean the problematic urban planning in a city using the corresponding taxi trajectories. Based on the traffic flow represented by taxi trajectories, the technology for detecting anomalies in urban areas has been reported in [9].

The work presented in this paper is also a step towards urban computing. Different from the above-mentioned research, however, we focus on the discovery of functional regions in a city, which we have never seen before in this research theme.

Discovery of functional regions: Functional regions [1] have been studied in traditional fields of GIS and urban planning for years, as the discovery of them can benefit policy making, resource allocation, and the related research. [7] gives a good survey on the related literatures which are mainly based on clustering algorithms. Some algorithms classify regions in urban area based on remote-sensor data, as thoroughly compared in [15]. Other network-based clustering algorithms (e.g., spectral clustering), however, employ interaction data, such as the economic transactions and people's movement between regions.

Recently, a brunch of work aims to study the geographic distribution of some topic in terms of user-generated social media. For example, Yin et al. [17] study the distributions of some geographical topics (like beach, hiking, and sunset) in USA using geo-tagged photos acquired from Flickr. Pozdnoukhov et al. [11] explore the space-time structure of topical content from a large number of geo-tweets. The social media generated in a geo-region is still used as static features to feature a region. On the other hand, a few literatures have reported that human mobility can describe the functions of regions. For instance, Qi et al. [12] observe that the getting on/off amount of taxi passengers in a region can depict the social activity dynamics in the region.

Our work is different from the research mentioned above in the following aspects. First, to the best of our knowledge, our method is the first one that simultaneously considers static features (POIs) of a region and interactions (human mobility) between regions when identifying functional regions. Second, rather than directly using some clustering algorithm, we propose a topic-model-based solution which represents a region with a distribution of functions. The function distribution is more practical than a single function for a region. Moreover, it reduces the data sparseness problem before clustering regions. We justify the advantage of our method over just using clustering approach in the experiments.

6. CONCLUSION

This paper proposes a framework (titled DRoF) for discovering regions of different functions (e.g., educational areas, entertainment areas, and regions of historic interests) in a city using both human mobility and POIs. The discovered functional regions help people easily understand a complex metropolitan, benefiting a variety of applications, such as urban planning, location choosing for a business, advertisement casting, and social recommendations. We evaluated this framework with a two-year Beijing POI dataset (2010 and 2011) and GPS trajectory datasets generated by over 12,000 taxis in year 2010 and 2011. According to the extensive studies, DRoF outperforms two baselines solely using POIs or mobility data in effectively finding functional regions. We also compared the results of DRoF in 2010 with that of 2011, discovering the evolving of Beijing. In addition, by matching the discovered functional regions against Beijing land use planning (2002-2010), we do not only validate our framework but also find interesting results.

In the future, we will further study the effectiveness of our method changing over the scale of the data we use. At the same time, we are going to employ or add other mobility data sources, such as cell-tower traces and check-ins in location-based services.

7. REFERENCES

[1] J. Antikainen. The concept of functional urban area. *Findings of the Espon project*, 1(1), 2005.

[2] S. Bednarz et al. *Geography for Life: National Geography Standards, 1994*. National Geographic Society, PO Box 1640, Washington, DC 20013-1640., 1994.

[3] D. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 2011.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] R. Estkowski. No Steiner point subdivision simplification is NP-Complete. In *Proc. 10th Canadian Conf. Computational Geometry*. Citeseer, 1998.

[6] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani. An energy-efficient mobile recommender system. In *Proc. KDD '10*, pages 899–908, 2010.

[7] C. Karlsson. Clusters, functional regions and cluster policies. *JIBS and CESIS Electronic Working Paper Series (84)*, 2007.

[8] L. Lam, S. Lee, and C. Suen. Thinning methodologies-a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 14(9):869–885, 1992.

[9] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *Proc. KDD '11*, pages 1010–1018, 2011.

[10] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, pages 411–418, 2008.

[11] A. Pozdnoukhov and C. Kaiser. Space-time dynamics of topics in streaming text. In *Proc. LBSN '11*, pages 8:1–8:8, 2011.

[12] 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 *IEEE PERCOM Workshops*, pages 384–388, 2011.

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

[14] L. Shapiro and G. Stockman. *Computer Vision. 2001*. Prentice Hall, 2001.

[15] 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 *Proc COM.Geo '11*, pages 11:1–11:10, 2011.

[16] M. Wand and M. Jones. *Kernel smoothing*, volume 60. Chapman & Hall/CRC, 1995.

[17] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *Proc. WWW '11*, pages 247–256, 2011.

[18] J. Yuan, Y. Zheng, X. Xie, and G. Sun. Driving with knowledge from the physical world. In *Proc. KDD '11*, pages 316–324, 2011.

[19] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun. Where to find my next passenger. In *Proc. Ubicomp '11*, pages 109–118, 2011.

[20] Y. Zheng, Y. Liu, J. Yuan, and X. Xie. Urban computing with taxicabs. In *Proc. Ubicomp '11*, pages 89–98, 2011.

[21] Y. Zheng and X. Zhou. *Computing with spatial trajectories*. Springer-Verlag New York Inc, 2011.