Exploring the Urban Region-of-Interest through the Analysis of Online Map Search Queries

Ying Sun[1,2,3], Hengshu Zhu[3*], Fuzhen Zhuang[1,2*], Jingjing Gu[4], Qing He[1,2]
[1]Key Lab of IIP of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China. {sunying17g, zhuangfuzhen, heqing}@ict.ac.cn
[2]University of Chinese Academy of Sciences, Beijing 100049, China.
[3]Baidu Inc., Beijing, China. zhuhengshu@baidu.com
[4]Nanjing University of Aeronautics and Astronautics, Nanjing, China. gujingjing@nuaa.edu.cn

ABSTRACT

Urban Region-of-Interest (ROI) refers to the integrated urban areas with specific functionalities that attract people's attentions and activities, such as the recreational business districts, transportation hubs, and city landmarks. Indeed, at the macro level, ROI is one of the representatives for agglomeration economies, and plays an important role in urban business planning. At the micro level, ROI provides a useful venue for understanding the urban lives, demands and mobilities of people. However, due to the vague and diversified nature of ROI, it still lacks of quantitative ways to investigate ROIs in a holistic manner. To this end, in this paper we propose a systematic study on ROI analysis through mining the large-scale online map query logs, which provides a new data-driven research paradigm for ROI detection and profiling. Specifically, we first divide the urban area into small region grids, and calculate their PageRank value as visiting popularity based on the transition information extracted from map queries. Then, we propose a density-based clustering method for merging neighboring region grids with high popularity into integrated ROIs. After that, to further explore the profiles of different ROIs, we develop a spatial-temporal latent factor model URPTM (Urban Roi Profiling Topic Model) to identify the latent travel patterns and Point-of-Interest (POI) demands of ROI visitors. Finally, we implement extensive experiments to empirically evaluate our approaches based on the large-scale real-world data collected from Beijing. Indeed, by visualizing the results obtained from URPTM, we can successfully obtain many meaningful travel patterns and interesting discoveries on urban lives.

ACM Reference Format:
Ying Sun[1,2,3], Hengshu Zhu[3*], Fuzhen Zhuang[1,2*], Jingjing Gu[4], Qing He[1,2]. 2018. Exploring the Urban Region-of-Interest through the Analysis of Online Map Search Queries. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220009

*Fuzhen Zhuang and Hengshu Zhu are corresponding authors.

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3220009

1 INTRODUCTION

With the development of market economies, the phenomenon of urban area agglomeration becomes more and more prevalent, which explicitly affects the urban business activities and all of the linked components. As one of the representatives, urban Region-of-Interest (ROI) attracts the attentions and activities of citizens in their daily lives. Different from Point-of-Interest (POI), ROI refers to the integrated urban areas with specific functionalities, such as the recreational business districts, transportation hubs, and city landmarks. Indeed, investigating on these ROIs can benefit multiple aspects of urban business. Specifically, at the macro level, ROI is an indicator for measuring the prosperity of agglomeration economies, which can facilitate the urban business planning for city administrations. At the micro level, ROI provides a useful venue for business practitioners to understand the urban lives, daily demands and mobilities of people, and thus enables a wide range of location-based services, including business site selection and targeted advertising.

In the literatures, prolific research efforts have been devoted to urban related topics based on the newly available mobile big data [7, 16, 24, 26]. However, due to the vague and diversified nature of ROI, it still lacks of quantitative ways to investigate ROIs in a holistic manner. For example, compared with the planned urban districts, the emergence of ROIs usually depends on the internal supply-demand relationships in the urban business environments. Therefore, it is a non-trivial task to accurately determine the range and boundaries of ROIs. Moreover, an ROI usually contains diversified functionalities, which might attract different visitors with different travel demands.

To this end, in this paper we propose a systematic study on ROI analysis through mining the large-scale online map query logs [21], which provides a new data-driven research paradigm for ROI detection and profiling. Indeed, compared with traditional mobile data used in urban computing, like check-ins and taxi trajectories [27], the online map queries contain richer contextual information for describing the travels in the urban environments, which can be leveraged for

Table 1: A toy example of map search query logs.

User ID	Time	Query Origin	Query Destination
UID#123	2015/12/15 16:13	Current Location	China Unicom
UID#124	2015/10/21 23:07	Current Location	HZ mansion
UID#125	2015/12/20 15:12	Xidan Joy City	Forbidden City
UID#126	2015/10/23 02:29	Current Location	Tian An Men
UID#127	2015/10/21 21:44	Sanlitun Village	Babaoshan
UID#128	2015/10/24 14:15	Current Location	Silk Street
UID#129	2015/12/18 2:18	Current Location	Happy east area 9

achieving fine-grained ROI analysis. Specifically, we first divide the urban area into small region grids, and calculate their PageRank values as visiting popularity based on the transition information (i.e., origin-destination) extracted from map queries. Then, we propose a density-based clustering method for merging neighboring region grids with high popularity into integrated ROIs. After that, to further explore the profiles of different ROIs, we develop a spatial-temporal latent factor model URPTM (Urban Roi Profiling Topic Model) to identify the latent travel patterns and Point-of-Interest (POI) demands of ROI visitors. In particular, URPTM can reveal the answers for which and why an ROI attracts people to visit. Finally, we implement extensive experiments to empirically evaluate our approaches based on the large-scale real-world data collected from Beijing. Indeed, by visualizing the results obtained from URPTM, we can successfully obtain many meaningful travel patterns and interesting discoveries on urban lives.

2 DATA DESCRIPTION

In this study, we use two sets of real-world data collected from a major commercial online map provider in China, namely map search query logs and urban POIs.

Specifically, the map search query logs contain the historical records of urban route search from map users in Beijing. For example, Table 1 shows a toy example of the query logs, where each record consists of an anonymized user ID, detailed query time, as well as the origin and destination of queries. In particular, the detailed GPS coordinates of both origin and destination in the queries are included, and most of the locations can be linked to a specific POI. Figure 1(a) and 1(b) demonstrate the geographical distribution of origins and destinations in map queries respectively. From the figures we can observe that the query origins distribute evenly in the city, while the destinations tend to concentrate in some agglomerated regions, which naturally motivates us to detect ROIs. Moreover, the map queries can reflect the travel pattern of citizens. For example, Figure 2 shows the urban travel flows among districts in Beijing on workday morning (8:00am-12:00am) extracted from map queries. Intuitively, based on the travel flows, we can easily estimate the popularity of different urban regions according to basic visiting frequency or graph analysis approaches, such as PageRank.

In the urban POI data, besides the detailed location information, a number of tags are also provided for each POI. The tags represent the functional categories of POIs, such as coffee house, general hospital, and shopping mall. Figure 1(c) shows the geographical distribution of POIs appearing in the map queries, meanwhile, Figure 1(d) shows the word-cloud representation of the tag frequencies of corresponding POIs.

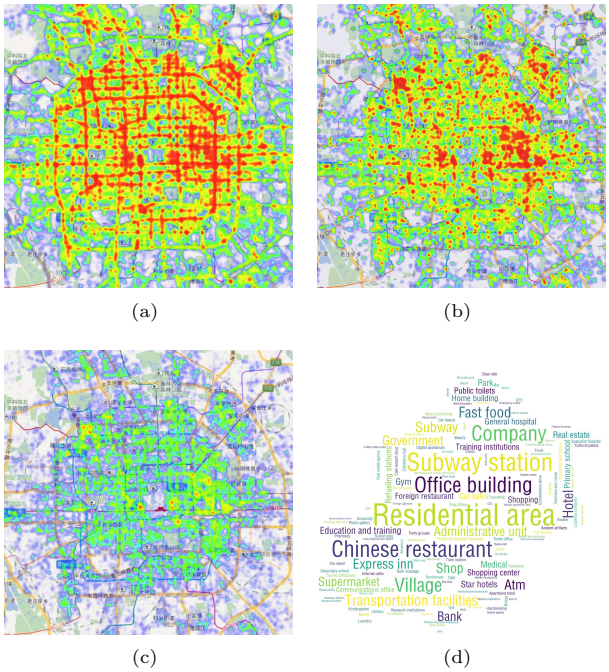


Figure 1: The geographical distributions of (a) origins, (b) destinations, and (c) POIs in map queries; (d) the frequency of POI tags in map queries.

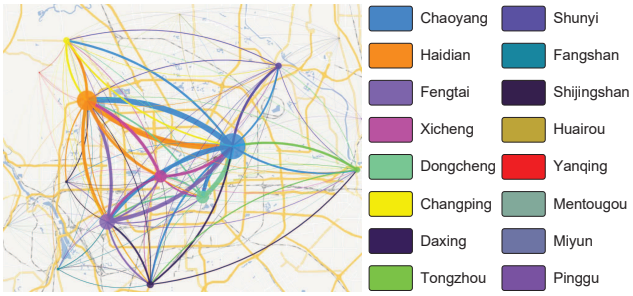


Figure 2: An example of urban travel flows on workday morning extracted from map queries, where the larger points have higher visiting frequencies.

3 PROBLEM FORMULATION

Based on the real-world data, the objective of this paper consists of two tasks, namely 1) *ROI Detection*, and 2) *ROI Profiling*, respectively. To be specific, for the task of *ROI Detection*, we aim to discover the integrated urban areas with high visiting popularity, based on the travel flow information extracted from map queries. For the task of *ROI Profiling*, we aim to model the spatial-temporal preferences of visitors

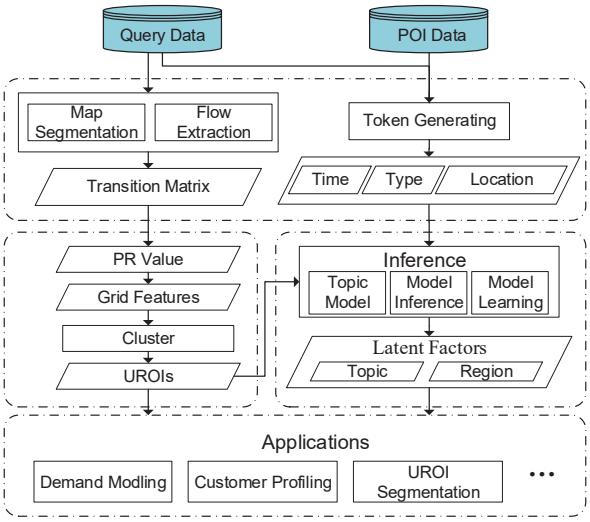
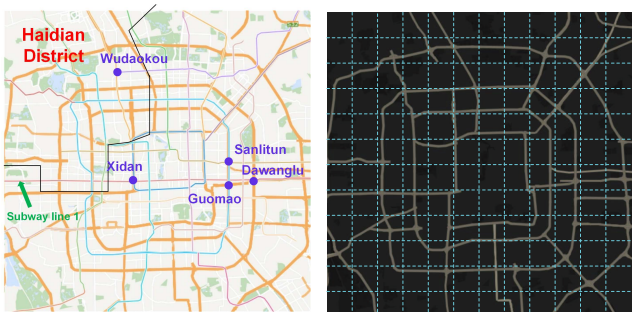


Figure 3: The framework overview of ROI analysis.



(a) Some landmarks (b) Grid-based partition

Figure 4: The urban area notations of Beijing.

for each ROI, which can reveal the answers for which and why an ROI attracts people to visit.

Figure 3 shows the overview of our ROI analysis framework. Specifically, in the first task, we first divide the urban area into small region grids. Then, by extracting the travel flows among these grids from query data, we form a transition matrix where each element denotes the query frequency between corresponding grids. After that, the PageRank algorithm is conducted on the matrix, in order to calculate the visiting popularity of each grid. With the grid popularity, we further cluster the grids with a density-based algorithm for detecting ROIs. In the second task, we design a spatial-temporal latent factor model URPTM for discovering the latent travel topics for ROI visitors. In the model, each ROI is regarded as a document, while time, origin and POI tags in the map queries, where the destination is located in the ROI, are regarded as words. After learning the model, we can get the spatial-temporal preferences of visitors for each ROI, which can be used for many applications, such as travel demand analysis and targeted ROI segmentation.

4 TECHNICAL DETAILS

In this section, we will introduce the technical details of our approaches for ROI analysis. To facilitate demonstration, we draw some landmarks in Figure 4(a).

4.1 ROI Detection Algorithm

Here, we introduce a density-based clustering method to effectively detect ROIs. Some notations are shown in Table 2. Specifically, the idea is to first split the urban area into small regions, and then merge them with clustering algorithm into ROIs. Note that, here we use the grid-based method for splitting urban regions [9], instead of the street-based segmentation [23], since the boundaries of ROIs are usually very vague and not consistent with the planned street blocks. Figure 4(b) shows an example of grid-based partition.

4.1.1 Calculating the Popularity of Region Grids. Since ROIs are the urban areas that attract people's attention and activities, we need an indicator to measure the visiting popularity, so that we can merge grids with higher visiting popularity into clusters. Although the visiting frequency is a straightforward indicator for popularity, it cannot capture the transition relationship between different regions. Intuitively, a region often queried by people from other popular regions should be more attractive. Moreover, when planning a trip to a POI, users may locate their query destinations at a nearby subway station, and then submit a query to a specific POI after they have arrived at places nearby. These regions are often traffic hubs and are not the real destinations of users. Therefore, in this paper we propose to use PageRank value for measuring the visiting popularity. With PageRank algorithm, the visiting popularity of traffic hubs, often acting as intermediate points with large flow in and large flow out, will converge to relatively small values. Specifically, based on the query data, we first construct an adjacency matrix, where each element denotes the query frequency between corresponding grids. Then PageRank is conducted based on the matrix, and the visiting popularity of each grid i is denoted as its PageRank value PR_i.

Meanwhile, we introduce an indicator **heat value** H_i to denote how the popularity of a grid i is influenced by grids around it. Specifically, we first define a heat kernel to calculate how g affects i, which is defined as: $h(i,g) = \exp(\frac{-dist(i,g)^2}{2\sigma})$, where $dist(i,g)$ denotes the Euclidean distance between center of grid i and grid g. Meanwhile, $h(i,g)$ will decay as $dist(i,g)$ increases. After that, we define the heat value of i as $H_i = \sum_{g \in \mathcal{N}_i} h(i,g) \times PR_g$, where $\mathcal{N}_i$ denotes the set of grids around i. Intuitively, the more popular the grids around i are, the higher value of H_i is.

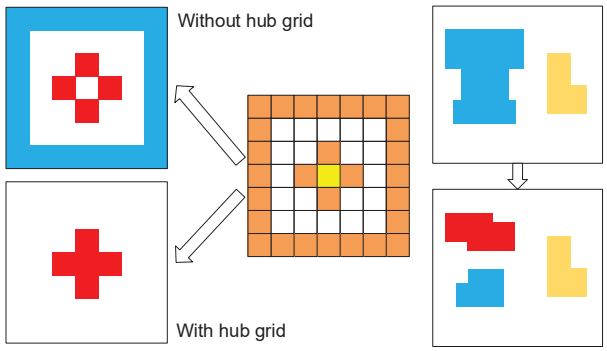
4.1.2 Density-based Clustering for ROI Detection. The general clustering process of our density-based clustering method is as follows. First, we start a new cluster with a new *core grid*, which is not ever clustered to existing clusters, and then

Table 2: Notations for our ROI detection algorithm.

Notation	Description
Popular grid	A grid i with $PR_i \geq PR_{th}$
Core grid	An active grids with at leat N_{th} popular grids in its neighborhood
H_i	Heat value of grid i
Hub Grid	A grid i whose $H_i \geq H_{th}$
Active grids	Popular grids and hub grids
Pop	The collection of popular grids
$Core$	The collection of core grids
PR_i	Visiting popularity of grid i
PR_{th}	Threshold for visiting popularity
H_{th}	Threshold for heat value
N_{th}	Threshold for neighborhood's size
$Msize$	Maximum size for ROIs
$Neigh(i)$	i's neighborhood (grids sharing an edge or a vertex with i)
Directly-reachable	Core grid i is directly-reachable to an active grid g if g is in $Neigh(i)$
Reachable	Grid p_1 is reachable to grid p_n if there exists a sequence $p_1, p_2, ..., p_n$ where p_i directly-reachable to p_{i+1}

add those core grids and *popular grids* that are *reachable* (refer to Table 2) by the current core grid to this clusters. Then, we repeat this process until all the core grids are clustered, and no new cluster is generated. Similar to the classic DBSCAN algorithm [3], our clustering algorithm is insensitive to the starting core grid, and we can obtain the same clustering results no matter which grid we start from. Specifically, we use breadth-first search (BFS) to implement the clustering algorithm, and the detailed procedure is shown in Algorithm 1.

Note that, we also consider the *hub grids* as core grids in the above process. Since at beginning, the whole urban area is divided into very small grids, and there may be some cases where popular grids in an ROI are separated by some unpopular grids. For example, a big shopping mall may contain popular grids with entrances and unpopular grids in its central area because users often locate their destinations at entrances when they are searching. Though do not always have large popularity, these grids usually have large heat values. We name a grid whose heat value higher than a threshold as a "hub grid". Furthermore, we give an example in Figure 5(a) to show how the hub grids affect the clustering process. The orange grids are popular grids and the yellow grid is an unpopular hub grid. Intuitively, the middle orange grids form an ROI. If we do not consider the yellow grid, to detect ROIs of this shape, we should set N_{th} no larger than 2 (or there will be no core grid among the middle 4 grids). This will lead the result to be the upper graph in Figure 5(a), where we find a red ROI and a blue ROI. However, the outer blue ring of grids are obviously not dense enough to form an ROI. If we use the yellow hub grid as a core grid, then we can set N_{th} up to 4 to get denser resulting ROIs, as the bottom graph in Figure 5(a).



(a) Hub grid (b) Re-partition

Figure 5: Examples of ROI detection.

In the real-world applications, visiting popularity is not evenly distributed over the whole urban area. Take Beijing as an example in Figure 1(b), we can see that the eastern areas of Beijing are more frequently queried by users. Thus grids in these areas will have higher visiting popularity than other grids. If we use the same thresholds PR_{th} to determine the *active grids* (refer to Table 2) in the whole urban area, the clustering algorithm will be failed to obtain reasonable results. A small value of PR_{th} may detect too many popular grids in the eastern areas and thus leads to very big clusters, while a large value of PR_{th} will lead to the missing of many clusters, e.g., some ROIs in the south of Beijing may be ignored. Intuitively, we should set different values of PR_{th} for different areas to detect all ROIs. Similar problem exists on H_{th}. To address this problem, we propose to limit the size of the clusters to automatically determine the thresholds. That is, if we find the size of a cluster output by Algorithm 1 is too large, i.e., number of core grids larger than the threshold $Msize$, then the value of threshold PR_{th} and H_{th} is increased for this area. The example of re-partition process is shown in Figure 5(b), the blue one will be re-partitioned into two smaller clusters and finally three ROIs are obtained. The completed algorithm for ROI detection with detailed adaptively re-partition procedure is shown in Algorithm 2. Note that we also increase $Msize$ as the procedure goes because intuitively an ROI of higher popularity may have a larger size.

4.2 ROI Profiling Model

Here, we propose to further explore the profiles of different ROIs, and develop a spatial-temporal latent factor model named URPTM (Urban Roi Profiling Topic Model) to identify the latent travel patterns of ROI visitors. Table 3 illustrates some notations and concepts of UPRTM.

4.2.1 ROI Profiling Topic Model. First, we would like to introduce how to adapt our ROI profiling problem to topic modeling setting, which considers the spatial and temporal information. Specifically, each ROI is regarded as a document d, and the word token is formed like (e, t, w), extracted

Algorithm 1 Density-based Region Grid Clustering

Require: Starting Core grid s, Pop, $Core$
Ensure: $Rset$: resulted cluster

function CLUSTER($s, Pop, Core$)
  $Rset \leftarrow \emptyset$
  $Q \leftarrow \{s\}$
  while $|Q| > 0$ **do**
    $u \leftarrow$ pop the first element in Q
    $Rset \leftarrow Rset \cup \{u\}$
    for $v \in Neigh(i) - (Rset \cup Q)$ **do**
      if $v \in Core$ **then**
        $Q \leftarrow Q \cup \{v\}$
      else
        if $v \in Pop$ **then**
          $Rset \leftarrow Rset \cup \{v\}$
  return $Rset$

Algorithm 2 ROI Detection

Require: G: grids,PR_{th}, $Msize$, H_{th}: the thresholds, I_1, I_2, I_3: the increasing values
Ensure: ROIs

function DISCOVER($G, PR_{th}, Msize, H_{th}$)
  $ROIs \leftarrow \emptyset$
  $V \leftarrow \emptyset$
  get $Core, Pop$
  $Q \leftarrow Core$
  while $|Q| > 0$ **do**
    for $g \in Q$ **do**
      if $g \in V$ **then**
        Continue
      $R \leftarrow$ CLUSTER($g, Pop, Core$)
      if $|Core \cap R| <= Msize$ **then**
        $V \leftarrow V \cup R$
        $ROIs \leftarrow ROIs \cup \{R\}$
    increase $PR_{th}, H_{th}, Msize$ with I_1, I_2, I_3 respectively
    get $Core, Pop$
    $Q \leftarrow Core - V$
  return $ROIs$

from the query data and the POI data, where e denotes the category of searched POI, w denotes the GPS coordinates of origin location, and t denotes the time stamp of the query. Meanwhile, we divide the time into several time slots so that t can be modeled as a discrete value.

We assume there are K topics reflecting users' travel patterns, intuitively, in each topic visitors from specific places would like to query for specific POI categories at specific time. So we model these topics to uncover the relationship between POI categories, time slots, query origins and ROIs shown in query data.

While POI categories and time slots can be associated with the topic easily with multinomial distributions. Since the query origins are spatial data, how to integrate non-discrete spatial information into our model becomes a big challenge. Intuitively, visitors from the same region tend to have similar travel demand. For example, in some areas people will show more interest in shopping because it is convenient for them to go to shopping centers. To this end, we creatively propose to

Table 3: Notations used in URPTM.

Notation	Description
R	The number of regions
K	The number of topics
D	The number of documents
N_d	The number of word tokens in document d
θ_d	The distribution of topics specific to document d
ϕ_z	The distribution of regions specific to topic z
η_z	The distribution of POI categories specific to topic z
ψ_z	The distribution of time slot specific to topic z
$z_{(d,i)}$	The topic assigned to token i in document d
$r_{(d,i)}$	The region assigned to token i in document d
$w_{(d,i)}$	The query location of token i in document d
$t_{(d,i)}$	The time slot of token i in document d
$e_{(d,i)}$	The POI type of token i in document d
τ_r	The Gaussian parameters of region r. $\tau_r = (\mu_r, \Sigma_r)$
α	The hyperparameter of Dirichlet prior on θ_d
β^1	The hyperparameter of Dirichlet prior on ϕ_z
β^2	The hyperparameter of Dirichlet prior on η_z
β^3	The hyperparameter of Dirichlet prior on ψ_z
γ	The hyperparameter of NIW prior on τ_r.$\gamma = (\mu_0, \Sigma_0, \Lambda_0, \nu_0, \kappa_0)$

model query origins with **Gaussian distributions**. Specifically, we assume there are finite number of topic regions in the urban area, and each topic region r is assigned with a Gaussian distribution with parameters $\tau_r = (\mu_r, \Sigma_r)$.

In URPTM, we assume that each document d is assigned with a multinomial distribution θ_d with respect to topics. And each topic z has three multinomial distributions η_z, ψ_z and ϕ_z with respect to e, t and r respectively. When generating a word token i in document d, a topic $z_{(d,i)}$ is first chosen according to θ_d. Then $e_{(d,i)}$, $t_{(d,i)}$ and $r_{(d,i)}$ are generated from the multinomial distributions. Finally, the query origin w is generated from the Gaussian distribution of $r_{(d,i)}$.

The graphical model of URPTM is shown in Figure 6, and the generative process is shown as follows:

1. For each demand region r:
   a. Draw $\tau_r = (\mu_r, \Sigma_r) \sim NIW(\gamma)$.
2. For each topic z:
   a. Draw $\phi_z \sim Dir(\beta^1)$.
   b. Draw $\eta_z \sim Dir(\beta^2)$.
   c. Draw $\psi_z \sim Dir(\beta^3)$.
3. For each document d:
   a. Draw $\theta_d \sim Dir(\alpha)$.
4. For each token i in document d:
   a. Draw $z_{(d,i)} \sim Multi(\theta_d)$.
   b. Draw $e_{(d,i)} \sim Multi(\eta_{z_{di}})$.
   c. Draw $t_{(d,i)} \sim Multi(\psi_{z_{(d,i)}})$.
   d. Draw $r_{(d,i)} \sim Multi(\phi_{z_{(d,i)}})$.
   e. Draw $w_{(d,i)} \sim N(\tau_{r_{(d,i)}})$.

4.2.2 Model Inference. For learning our model URPTM, the Gibbs sampling [5] is adopted to infer all the parameters. Specifically, for updating z, we have

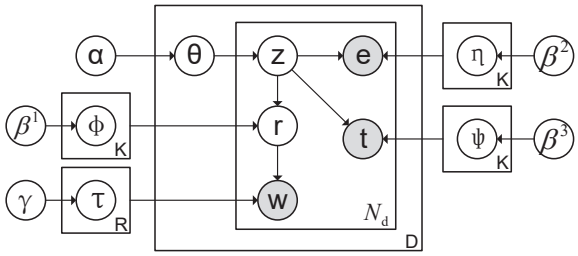


Figure 6: Urban Roi Profiling Topic Model.

$$\begin{aligned} p(z_{(d,i)} = k|z_{\neg(d,i)}, e, t, r) \propto\ & p(z_{(d,i)} = k|z_{\neg(d,i)}) \\ & \cdot p(e_{(d,i)}|z_{(d,i)} = k, z_{\neg(d,i)}, e_{\neg(d,i)}) \\ & \cdot p(t_{(d,i)}|z_{(d,i)} = k, z_{\neg(d,i)}, t_{\neg(d,i)}) \\ & \cdot p(r_{(d,i)}|z_{(d,i)} = k, z_{\neg(d,i)}, r_{\neg(d,i)}), \end{aligned} \tag{1}$$

where $z_{(d,i)} = k$ represents the assignment of the topic to the ith token in the dth document (referred as token (d,i)), $z_{\neg(d,i)}$ represents the assignment of the topic to all the tokens excluding token (d,i). Similarly, $r_{(d,i)}$ represents the assignment of the region for token (d,i) and $r_{\neg(d,i)}$ represents the assignments excluding token (d,i). The notations about e and t can be similarly inferred. To be specific, we have

$$p(z_{(d,i)} = k|z_{\neg(d,i)}) = \frac{O^k_{d,\neg(d,i)} + \alpha}{\sum_{k=1}^{K} O^k_{d,\neg(d,i)} + K\alpha}, \tag{2}$$

$$p(e_{(d,i)} = e|z_{(d,i)} = k, z_{\neg(d,i)}, e_{\neg(d,i)}) = \frac{O^e_{k,\neg(d,i)} + \beta^1}{\sum_{e=1}^{E} O^e_{k,\neg(d,i)} + E\beta^1}, \tag{3}$$

$$p(t_{(d,i)} = t|z_{(d,i)} = k, z_{\neg(d,i)}, t_{\neg(d,i)}) = \frac{O^t_{k,\neg(d,i)} + \beta^2}{\sum_{t=1}^{T} O^t_{k,\neg(d,i)} + T\beta^2}, \tag{4}$$

$$p(r_{(d,i)} = r|z_{(d,i)} = k, z_{\neg(d,i)}, r_{\neg(d,i)}) = \frac{O^r_{k,\neg(d,i)} + \beta^3}{\sum_{t=1}^{R} O^r_{k,\neg(d,i)} + R\beta^3}, \tag{5}$$

where $O^k_{d,\neg i}$ denotes the frequency that the kth topic assigned in document d excluding token (d,i). K denotes the number of topics. $O^e_{k,\neg(d,i)}$ denotes the frequency that the kth topic assigned to a token with eth POI type in the corpus excluding token (d,i). E denotes the number of POI types. $O^t_{k,\neg(d,i)}$ denotes the frequency that the kth topic assigned to a token with tth time stamp in the corpus excluding token (d,i). T denotes the number of time stamps. $O^r_{k,\neg(d,i)}$ denotes the frequency that the kth topic assigned to a token with rth region in the corpus excluding token (d,i). R denotes the number of regions. For region r, we have

$$p(r_{(d,i)} = r|z, e, t, w, r_{\neg(d,i)}, w) = p(r_{(d,i)} = r|z, r_{\neg(d,i)}) \tag{6}$$

$$\propto p(r_{(d,i)}|z_{(d,i)} = k, z_{\neg(d,i)}, r_{\neg(d,i)}) \cdot p(w_{(d,i)}|r_{(d,i)} = r, w_{\neg(d,i)}), \tag{7}$$

$$p(w_{(d,i)}|r_{(d,i)} = r, w_{\neg(d,i)}) = p(w_{(d,i)}|W_{(d,i)}), \tag{8}$$

where $W_{(d,i)}$ denotes the set of all the points in tokens excluding (d,i) which is assigned with the same region r as token (d,i). The posterior distribution $p(w|W)$ follows a multivariate Student-T distribution [18] whose parameters are similar with the distribution in [19].

5 EXPERIMENTS

In this section, we evaluate our framework on map search query data, including the evaluation of ROI detection and profiling, and a case study on ROI Segmentation.

5.1 Experimental Setup

We collected the online map search query data of Beijing in three weeks. After filtering the failed queries and queries outside of Beijing, and deleting duplicate queries with the same destination submitted by the same user in a short time (ten minutes is used in this paper), we obtained more than 1,500,000 queries. On the other hand, we acquired the POI categories of the destinations by searching the keywords and destination from the online map API, and obtained about 120,000 POIs with 158 kinds of tags. For this POI dataset, we also conducted some preprocessing to delete the POIs with no tags and merge some tags, e.g., "East Gate of Tsinghua University" and "Tsinghua University" were merged into "Tsinghua University".

We divided the urban area of Beijing into small region grids with size $L \times L\ m^2$ according to the Geographic Information System (GIS) coordinate. Intuitively, the different setting of L will lead to different accuracy in ROI detection. For conducting fine-grained ROI analysis, we empirically set $L = 200$ in our experiments.

In ROI detection, the hyper-parameters can be adjusted according to different application requirements. For instance, if we want to obtain denser ROI distribution, we can increase N_{th}. Similarly, if we increase PR_{th} and H_{th}, ROIs with higher visiting popularity would be obtained, and vice versa. In our experiments, we empirically set $\sigma = 0.2$ when computing heat values and $\mathcal{N}_i$ as the 8 grids sharing vertices or edges with i. Moreover, we set $N_{th} = 3$, the initial value for $Msize$, PR_{th}, H_{th} as 3, 0.00003, 0.0003 and let them increase by 0.2, 0.0000005, 0.0000005 each round. When generating tokens, we used the query data with destination in detected ROIs and split time into 48 time slots denoting 24 hours on workdays and weekends. For URPTM, we assume there are 20 topics and 200 topic regions. We choose the priors $\alpha = 2.5$, $\beta^1 = 0.25$, $\beta^2 = 1$, $\beta^3 = 0.3$ with preliminary experiments, and set the supreme parameters in T-distribution as $\mu_0 = \frac{\sum_{d,i} w_{(d,i)}}{N}, \Lambda_0 = \begin{bmatrix} 0.001 & 0.0005 \\ 0.0005 & 0.001 \end{bmatrix}, \kappa_0 = 2, \nu_0 = 2$, where N denotes the number of tokens. The maximum number of iterations in Gibbs sampling is set to 1,000.

5.2 Evaluation on ROI Detection

In our ROI detection algorithm, we use PageRank value to denote the visiting popularity of each grid. To validate the advantage of PageRank, we choose another popularity criterion by using query frequency as baseline. Some examples of comparison results are shown in Figure 7, where blue and red grids are top-500 popular grids of the same regions in Beijing, generated with frequency and PageRank respectively.

From the figures we can observe that some grids with high visiting frequency are not regarded as popular based on

(a) Jiaotong University (b) Baofusi Bridge

Figure 7: Popular grids obtained by different criteria (Blue: Frequency, Red:PageRank).

PageRank values. For example, in Figure 7(a), the blue grid is located near a bus stop, while the red grid is located in Beijing Jiaotong University. Although there is also a building in the blue grid, it is actually inside the campus. Similarly, in Figure 7(b), the red grid contains more buildings, restaurant and swimming pool compared with the blue grid. In both figures, the red grids are more possible to be the final destinations of visitors.

Figure 8 shows the total 188 ROIs detected by our algorithm in Beijing, where different colors are used to identify ROIs. In general, northern and eastern parts of Beijing have more ROIs than the southern and western parts. Furthermore, there are also some well-known ROIs, such as Guomao CBD[1], and Zhongguancun[2]. Meanwhile, we can find that different ROIs have different composition of POIs, e.g., office buildings and shopping malls in Zhongguancun, shopping malls in Xidan[3], restaurants in Wudaokou[4] and office buildings in Guomao CBD.

5.3 Evaluation on ROI Profiling

In this subsection, we evaluate the performance of our model URPTM in ROI profiling, including topic modeling and visitor inference for ROIs.

5.3.1 Topic Modeling for ROIs. In Table 4, we demonstrate the detailed information of 5 randomly selected topics learned by URPTM. Specifically, the first row shows the spatial distribution of the topic, where we use the heat map to represent the probability of visitors in different locations be attracted by the topic, i.e., $p(w|z) = \sum_r p(w|r)p(r|z)$. The second row shows the major POI categories contained in the topics. The third row shows temporal distribution of the topic, where the x-axis contains 48 time slots (i.e., 0:00am-23:00pm weekday, 0:00am-23:00pm weekend). Note that since the number of time slots in weekdays are 2.5 times of that in weekends, for facilitating the demonstration, we re-scaled the range by dividing the values of weekdays by 2.5.

From the results, we can obtain intuitive understanding about the topics. Take topic 18 as an example, we can find that it is about the travel pattern of "Going to Work", occurring mainly on workday mornings. From the spatial distribution, we can find that this topic attracts more people from the

[1]https://en.wikipedia.org/wiki/Guomao,_Beijing
[2]https://en.wikipedia.org/wiki/Zhongguancun
[3]https://en.wikipedia.org/wiki/Xidan
[4]https://en.wikipedia.org/wiki/Wudaokou

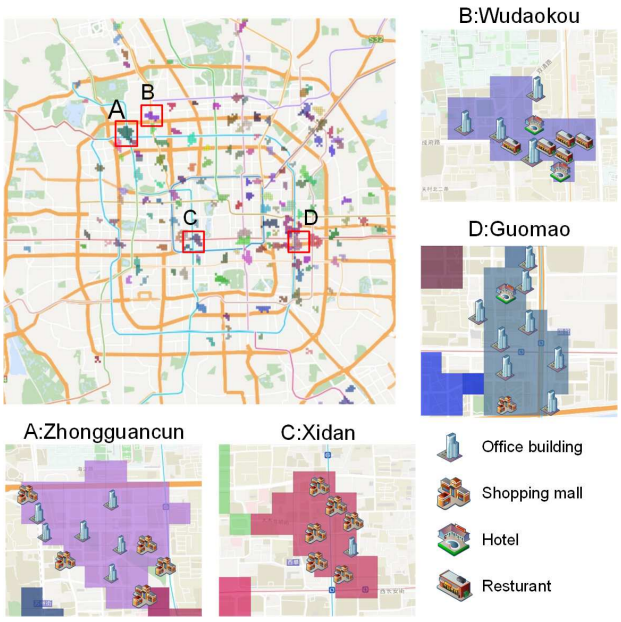


Figure 8: ROIs detected in Beijing.

northeast of Beijing (i.e., one of the major residential areas in Beijing). Indeed, from these topics we can understand the urban lives of people.

EXAMPLE 1. *On a weekday morning, patients go to hospital for registering, some of them even set out at 5:00 am (based on Topic 17). Several hours later, people will go to work from home (based on Topic 18). After working for all day, they will go to some restaurants with family for dinner in a shopping mall (based on Topic 7), while most of them would like to go shopping on weekends (based on Topic 15). For the college students, their travel patterns have a vague boundary between workdays and weekends (based on Topic 14).*

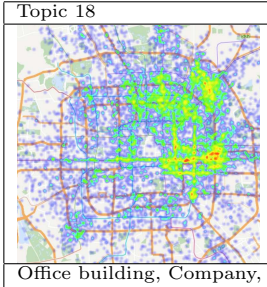
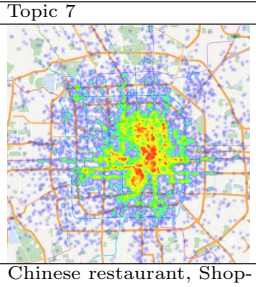
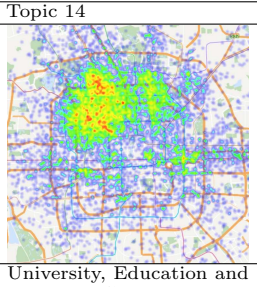
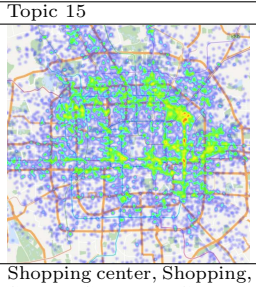
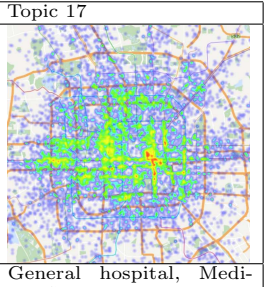
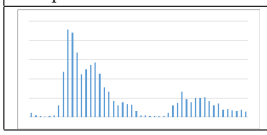
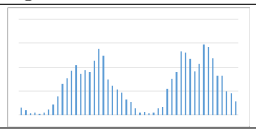
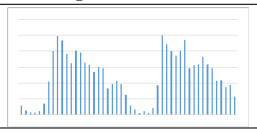
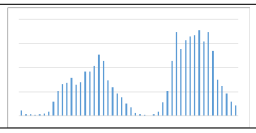
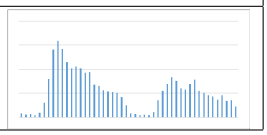
Meanwhile, these topics can provide guidance for the urban business planning, such as business site selection. For example, the office buildings can be built according to the spatial distribution of topic 18, and hospitals can be built according to topic 7.

5.3.2 Global Visitor Inference for ROIs. Knowing the origins of visitors for different ROIs can benefit many business applications, such as targeted advertising. Intuitively, we can obtain this distribution by counting the origins in the map queries. However, even with the large-scale query data, the observed origins of ROI visitors are usually very sparse. Indeed, our model URPTM can be used for globally interring the visitor distributions of ROIs. Specifically, we can use the probability $p(w|d) = \sum_r p(w|r)p(r|d)$ for estimating the origins of visitors in ROI d.

Particularly, we take three well-known ROIs, i.e., Zhongguancun, Xidan and Sanlitun[5], for model evaluation. Figure 9(a), 9(b) and 9(c) show the exact query origins to these

[5]https://en.wikipedia.org/wiki/Sanlitun

Table 4: The details of some topics learned by URPTM.

Topic 18	Topic 7	Topic 14	Topic 15	Topic 17
				
Office building, Company, Real estate, Park, Bank, Administrative unit, Residential area, Public prosecution agencies, Company, Corporation	Chinese restaurant, Shopping, Theater, Star Hotels, Office building, Hotel, Express Inn, Shop, Food, Fast food, Shopping center, Foreign restaurant	University, Education and training, Chinese restaurant, Subway station, Residential area, Office building, Refueling stations, Training institutions	Shopping center, Shopping, Shop, Cinema, Chinese restaurant, Express inn, Snack shop, Dessert cake, Entertainment	General hospital, Medical, Subway station, Hotel, Cultural palace, Primary school, Express inn, Snack shop
				

ROIs in our dataset. We can observe that though they are well-known business districts, only local visitors can be identified, visitors from other districts are very orphan words.

Figure 9(d), 9(e) and 9(f) shows the inferred distribution of visitors by URPTM, which are more reasonable and comprehensive. Take Xidan as an example, we find that it can attract visitors along subway Line 1 (i.e., the east-west straight line on the middle of Beijing), which is reasonable as Xidan locates on Line 1, and it is convenient for these visitors to go there. Therefore, a shop at Xidan can make advertisement on subway line 1 to attract more customers. Moreover, Xidan can also attract visitors from the north of Beijing, where many universities are located.

5.3.3 Topic-Aware Visitor Inference for ROIs. Besides the global visitor inference for ROIs, URPTM also can be used for inferring the origins of ROI visitors with respect to different topics. For example, where are the origins of people who visit Xidan for dinner? Specifically, for an ROI d and topic z, we can obtain the origin distributions by computing $p(w|d,z) = \sum_r p(w|r) \sum_z p(r|z)p(z|d)$.

Particularly, since the classic Latent Dirichlet Allocation (LDA) [2] can also assign tokens with topics, here we use it as baseline for comparison. Similar to URPTM, we consider ROIs as documents and each queried POI category as a token. At the step of visualization, the locations are drew based on the origins of most possibly queried POI categories in ROI topics with heat map.

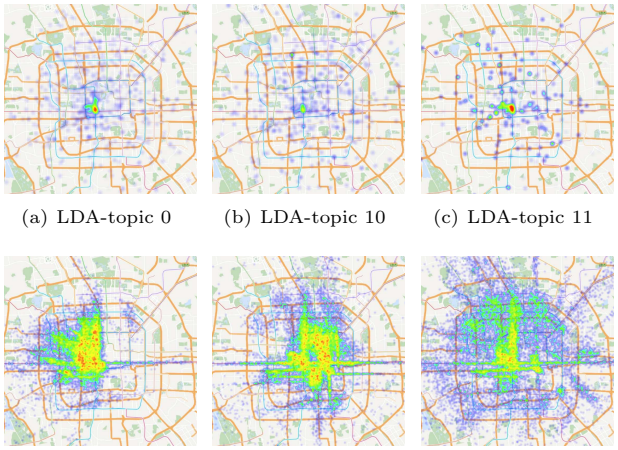
We present the results of different models for Xidan with 3 major learned topics in Figure 10. We can find that the results obtained by URPTM are more reasonable, and the origins of visitors for different topics are more clearly identified. For example, Figure 10(e) shows the visitors for topic 7 learned by URPTM, where the POIs are mainly about "Restaurant". Figure 10(f) shows the visitors for topic 15, where the POIs are mainly about "Shopping". Compared with topic 7, topic 15 attracts less visitors from the east of Beijing. Indeed, many shopping centers are located in the east of Beijing and thus Xidan is not so attractive for visitors there in terms of shopping. Meanwhile, Topic 7 is not as attractive as topic 15 to visitors in Haidian district[6], since people there prefer Wudaokou for restaurants. Actually, based on the visitor inference of different topics, we can also obtain the topic-aware decomposition of visitors in different ROIs. Figure 11 shows visitors to Guomao as an example. The visitors to an ROI is composed of visitors with different topics.

(a) Frequency-ZGC (b) Frequency-XD (c) Frequency-SLT

(d) URPTM-ZGC (e) URPTM-XD (f) URPTM-SLT

Figure 9: Visitor inference for ROIs. ZGC: Zhongguancun, XD: Xidan, SLT: Sanlitun.

[6] https://en.wikipedia.org/wiki/Haidian_District



(a) LDA-topic 0 (b) LDA-topic 10 (c) LDA-topic 11

(d) URPTM-topic 0 (e) URPTM-topic 7 (f) URPTM-topic 15

Figure 10: Topic-aware visitor inference for Xidan.

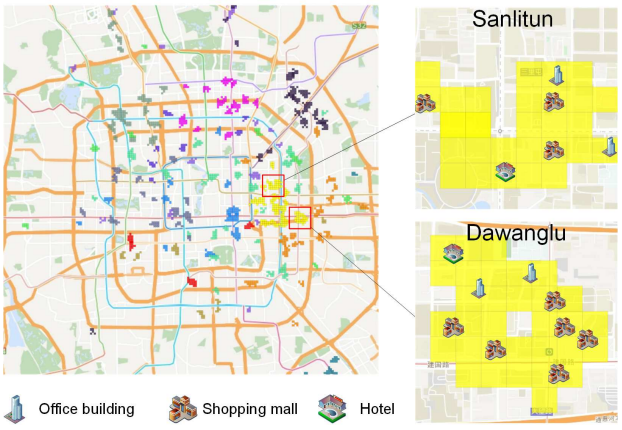


Figure 12: The ROI segmentation results.

5.4 Case Study on ROI Segmentation

In this subsection, we introduce a case study on ROI segmentation, which aims to segment ROIs into different clusters with respect to their profiles. Indeed, ROI segmentation plays an important role for urban business planning. For example, if two ROIs have similar topics, it is reasonable to carry out similar promotions and advertising campaigns for them.

Specifically, based on the results from URPTM, we can segment ROIs into different clusters based on their topic distribution. Figure 12 shows the ROI segmentation results by K-Means algorithm, where 15 clusters are generated and ROIs in the same cluster have same color. Generally, by manually inspecting the results, we think the segmentation results are reasonable. For example, in both Sanlitun and Dawanglu, there are many shopping malls and hotels, and their visitors are majorally from the east of Beijing. Therefore, if someone wants to open a cafe at Sanlitun but for some reason has to change the location, Dawanglu is another good choice, since they all belong to the same cluster and thus has similar visitors.

6 RELATED WORK

Generally, the related works of this paper can be grouped into two categories, namely *Urban POI Analysis* and *Topic Model Applications*.

Urban POI Analysis. In the past years, POI analysis has attracted wide attention in urban computing related domains [12, 29], and thus a number of relevant studies have been introduced for various applications like POI recommendation [10, 11], demand modeling and urban function discovery [13, 22, 25]. For example, Liu *et al.* [14] proposed a geographical probabilistic model for POI recommendations, and Li *et al.* [10] designed a novel POI recommender system based on the check-ins of friends. Furthermore, some researchers proposed to leverage the POIs for urban region analysis. For example, Yuan *et al.* [23, 24] used POIs and the GPS trace trajectory of cabs to discover the function regions in urban area. Liu *et al.* [15] proposed a framework named RPDI to identify POI demands of different regions, where POI profiles, demographic data and human mobility data are modeled together. Recently, Hu *et al.* [8] introduced a method to extract and understand the urban areas-of-interest using POI-tagged photos. Different from the above studies, in this paper we introduced a systematic study on ROI analysis through mining the large-scale online map query logs.

Topic Model Applications. Topic models are among the most popular techniques for modeling discrete data, such as textual data. In the past decades, a large number of topic models have been proposed, such as Latent Dirichlet Allocation (LDA) [2], Probabilistic Latent Semantic Indexing (PLSI) [6], Topic-Over-Time model (TOT) [20] and Dynamic Topic Model (DTM) [1]. Indeed, these topic models were applied in various applications over different domains. For example, Zhu *et al.* [28] proposed a time-aware topic model to track evolution of social emotions. Fu *et al.* [4] proposed a geographic topic model to extract mobility patterns for estate ranking prediction. For consumer behavior analysis, Luo *et al.* [17] introduced a solution for online-to-offline recommendations via topic model. Wang *et al.* [19] designed a topic model named T^{3}M for learning tactical patterns of soccer teams, which creatively uses the Gaussian priors for modeling location information. Inspired by the above works, in this paper we developed a spatial-temporal latent factor model URPTM to identify the latent travel patterns and POI demands of ROI visitors.

7 CONCLUSION

In this paper, we introduced a systematic study on ROI analysis through mining the large-scale online map query logs. Specifically, we first divided the urban area into small region grids, and calculated their Pagerank value as visiting popularity based on the transition information extracted from map queries. Then, we proposed a density-based clustering method

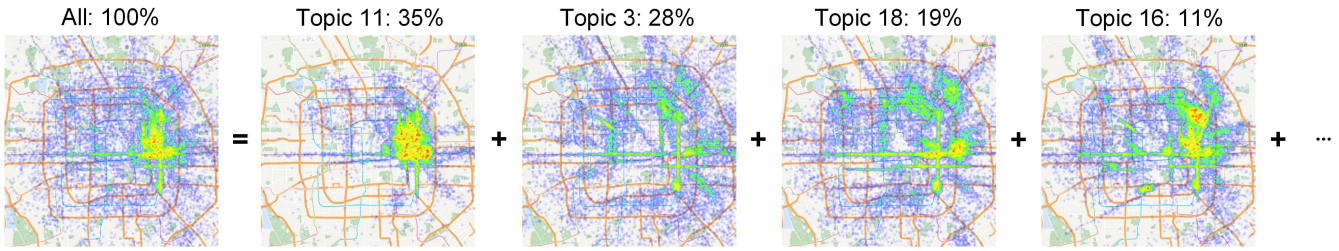


Figure 11: Topic-aware decomposition of visitors inference for Guomao.

for merging neighboring region grids into integrated ROIs. After that, to further explore the profiles of different ROIs, we developed a spatial-temporal latent factor model URPTM to identify the latent travel patterns and POI demands of ROI visitors. Finally, we conducted extensive experiments to empirically evaluate our framework based on the large-scale real-world data collected from Beijing. The experimental results clearly validated the effectiveness and interpretability of our approaches in terms of ROI detection and profiling.

ACKNOWLEDGMENTS

The research work is supported by the National Key R&D Program of China (2018YFB1004300), the National Natural Science Foundation of China under Grant Nos. 61773361, 61473273, 91546122, Guangdong provincial science and technology plan projects under Grant No. 2015B010109005 and the Youth Innovation Promotion Association CAS 2017146.

REFERENCES

[1] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of ICML*. 113–120.
[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* (2003), 993–1022.
[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Proceedings of ACM SIGKDD*. 226–231.
[4] Yanjie Fu, Guannan Liu, Spiros Papadimitriou, Hui Xiong, Yong Ge, Hengshu Zhu, and Chen Zhu. 2015. Real estate ranking via mixed land-use latent models. In *Proceedings of ACM SIGKDD*. 299–308.
[5] Gregor Heinrich. 2008. Parameter Estimation for Text Analysis. *Technical Report* (2008).
[6] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the UAI*. 289–296.
[7] Liang Hong, Yu Zheng, Duncan Yung, Jingbo Shang, and Lei Zou. 2015. Detecting urban black holes based on human mobility data. In *Proceedings of the 23rd SIGSPATIAL*. 35:1–35:10.
[8] Yingjie Hu, Song Gao, Krzysztof Janowicz, Bailang Yu, Wenwen Li, and Sathya Prasad. 2015. Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems* (2015), 240–254.
[9] John Krumm and Eric Horvitz. 2007. Predestination: Where do you want to go today? *Computer* (2007).
[10] Huayu Li, Yong Ge, Richang Hong, and Hengshu Zhu. 2016. Point-of-interest recommendations: Learning potential check-ins from friends. In *Proceedings of ACM SIGKDD*. 975–984.
[11] Xutao Li, Gao Cong, Xiao-Li Li, Tuan-Anh Nguyen Pham, and Shonali Krishnaswamy. 2015. Rank-geofm: A ranking based geographical factorization method for point of interest recommendation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 433–442.
[12] Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. 2014. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 831–840.
[13] Bin Liu and Hui Xiong. 2013. Point-of-interest recommendation in location based social networks with topic and location awareness. In *Proceedings of SIAM SDM*. 396–404.
[14] Bin Liu, Hui Xiong, Spiros Papadimitriou, Yanjie Fu, and Zijun Yao. 2015. A general geographical probabilistic factor model for point of interest recommendation. *IEEE TKDE* (2015), 1167–1179.
[15] Yanchi Liu, Chuanren Liu, Xinjiang Lu, Mingfei Teng, Hengshu Zhu, and Hui Xiong. 2017. Point-of-Interest Demand Modeling with Human Mobility Patterns. In *Proceedings of ACM SIGKDD*. 947–955.
[16] Ye Liu, Yu Zheng, Yuxuan Liang, Shuming Liu, and David S Rosenblum. 2016. Urban water quality prediction based on multi-task multi-view learning. In *international joint conference on artificial intelligence*. 2576–2581.
[17] Ping Luo, Su Yan, Zhiqiang Liu, Zhiyong Shen, Shengwen Yang, and Qing He. 2016. From online behaviors to offline retailing. In *Proceedings of ACM SIGKDD*. 175–184.
[18] Christian Robert. 2014. Machine learning, a probabilistic perspective. (2014).
[19] Qing Wang, Hengshu Zhu, Wei Hu, Zhiyong Shen, and Yuan Yao. 2015. Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications. In *Proceedings of ACM SIGKDD*. 2197–2206.
[20] Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of ACM SIGKDD*. 424–433.
[21] Xiangye Xiao, Qiong Luo, Zhisheng Li, Xing Xie, and Wei-Ying Ma. 2010. A large-scale study on map search logs. *ACM Transactions on the Web (TWEB)* 4, 3 (2010), 8.
[22] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wen-Ning Kuo, and Vincent S Tseng. 2012. Urban point-of-interest recommendation by mining user check-in behaviors. In *Proceedings of ACM SIGKDD*. 63–70.
[23] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of ACM SIGKDD*. 186–194.
[24] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. 2015. Discovering urban functional zones using latent activity trajectories. *IEEE TKDE* (2015), 712–725.
[25] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Time-aware point-of-interest recommendation. In *Proceedings of ACM SIGIR*. 363–372.
[26] Xianyuan Zhan, Yu Zheng, Xiuwen Yi, and Satish V Ukkusuri. 2017. Citywide traffic volume estimation using trajectory data. *IEEE TKDE* 29, 2 (2017), 272–285.
[27] Yu Zheng. 2015. Trajectory data mining: an overview. *ACM TIST* 6, 3 (2015), 29.
[28] Chen Zhu, Hengshu Zhu, Yong Ge, Enhong Chen, and Qi Liu. 2014. Tracking the evolution of social emotions: A time-aware topic modeling perspective. In *IEEE ICDM*. 697–706.
[29] Hengshu Zhu, Hui Xiong, Fangshuang Tang, Qi Liu, Yong Ge, Enhong Chen, and Yanjie Fu. 2016. Days on market: Measuring liquidity in real estate markets. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 393–402.