

Mining Shopping Patterns for Divergent Urban Regions by Incorporating Mobility Data*

Tianran Hu^{*1}, Ruihua Song^{†2}, Yingzi Wang^{†§3}, Xing Xie^{†4}, Jiebo Luo^{*5}

^{*}University of Rochester [†]Microsoft Research

[§]University of Science and Technology of China

{¹thu, ⁵luo}@cs.rochester.edu,

{²song.ruihua, ⁴xing.xie}@microsoft.com, ³yingzi@mail.ustc.edu.cn

ABSTRACT

What people buy is an important aspect or view of lifestyles. Studying people's shopping patterns in different urban regions can not only provide valuable information for various commercial opportunities, but also enable a better understanding about urban infrastructure and urban lifestyle. In this paper, we aim to predict city-wide shopping patterns. This is a challenging task due to the sparsity of the available data – over 60% of the city regions are unknown for their shopping records. To address this problem, we incorporate another important view of human lifestyles, namely mobility patterns. With information on “where people go”, we infer “what people buy”. Moreover, to model the relations between regions, we exploit spatial interactions in our method. To that end, Collective Matrix Factorization (CMF) with an interaction regularization model is applied to fuse the data from multiple views or sources. Our experimental results have shown that our model outperforms the baseline methods on two standard metrics. Our prediction results on multiple shopping patterns reveal the divergent demands in different urban regions, and thus reflect key functional characteristics of a city. Furthermore, we are able to extract the connection between the two views of lifestyles, and achieve a better or novel understanding of urban lifestyles.

Keywords

Shopping Patterns, Mobility Patterns, Urban Computing, Multi-view Lifestyles

1. INTRODUCTION

Studies on land use suggest a variety of types of city regions, such as residential, industrial and others [27, 3]. Such divergence of city region functionality [29] naturally leads to the various shopping patterns in different regions. Daily supplies, for example, are usually needed in residential quarters, while office appliances are more popular in business areas. In this paper, we aim to predict

*This work was conducted while the first author was doing internship at Microsoft Research Asia

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM'16, October 24-28, 2016, Indianapolis, IN, USA

© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...\$15.00

DOI: <http://dx.doi.org/10.1145/2983323.2983803>

shopping patterns in different urban regions. Through the prediction, we are able to read the various demands of these regions, gain knowledge of the infrastructures of urban areas, and more importantly, acquire a better understanding of human lifestyle geographically. Furthermore, this study could be used to support location-based marketing, such as mobile advertising, outdoor exhibition stands, and so on.

Market Basket Analysis shows that consumers usually have demands for a group of products [4], and people's demands are highly related to their lives [9]. For example, an IT practitioner may have interests in digital products, computer games, as well as technical books. Motivated by this observation, we consider shopping patterns as combinations of inherently related types of products. Our goal in this study is to predict the levels of shopping patterns in different city regions. Compared with predicting the levels of single types of products, there are two main benefits to doing so. First, since products are actually correlated, the prediction of a group of related products provides more valuable information for item recommendation and commercial planning. Second, since shopping patterns are more related to the lifestyles of people in a region than single categories of products, predicting shopping patterns can give us a more comprehensive understanding of urban areas.

In this study, we collect geolocated logs of browsing products online from Beijing, and extract shopping patterns for the regions in the city. Although the visiting volumes to product webpages are not the exact product trading volumes, they indicate the demands for different products. Therefore, the browsing records can be regarded as reasonable proxy of consumer shopping behaviors. According to the study on human mobility [28], the probability of a person moving a certain distance drops dramatically as the distance exceeds 1km in urban area. Therefore, we define a region as a 1km by 1km grid along the latitude and longitude lines in a city. We formalize region-based shopping patterns as a matrix, and an element in the matrix denotes the level of a shopping pattern in a region. Our task is to complete the missing values in the matrix. The matrix suffers from a severe sparsity problem – over 60% of rows are empty. In other words, we do not have any shopping information of more than 60% regions in the city. To address this problem, we seek help from other types of lifestyle information.

Lifestyle is defined as “the way in which a person or a group of people live” in Merriam-Webster Dictionary¹. There are many aspects or views to describe how people live [24, 30], including shopping patterns and mobility patterns. For example, a teenage student may have a certain shopping pattern, such as books, digital products, and so on. Similar to shopping patterns, we consider a mobility pattern as a group of inherently related POI categories. We can also see a certain mobility pattern of the student, such as

¹<http://www.merriam-webster.com/dictionary/lifestyle/>

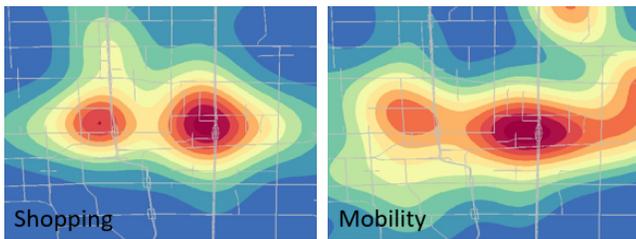


Figure 1: The illustration of the geographical distributions of a shopping pattern (left) and a mobility pattern (right) of the same region on a local map in Beijing.

dormitory, library, and so on. Multiple views of a lifestyle are actually connected, and this connection is reflected on their geographical distributions. Figure 1 shows the geographical distributions of a shopping pattern extracted from the browsing logs and a mobility pattern extracted from check-in data in a region. The top three most browsed product categories of the illustrated shopping pattern (left) are Laptops, Office Appliances & Stationery, and Tablets. The top three most checked in POI categories of the mobility pattern (right) are Offices, Convenience Stores, and Commute. We observe that the geographical distributions of these two patterns are quite similar, and they both reach the highest levels in the same block. In other words, many people check in at offices (mobility) in this block, meanwhile, many people show high demands for office appliances (shopping) in the same block. The geographical connection of these two views of the lifestyle inspires this work: connecting shopping and mobility patterns, and utilizing the mobility patterns in a region to predict the shopping patterns there. To achieve this, we learn region mobility patterns from check-in data of a popular Location Based Social Network (LNSB) in China, and employ a Collective Matrix Factorization (CMF) [21] model to fuse these two views.

However, considering the lifestyles of regions in a city separately is inappropriate. Regions in a city are never isolated, as they are connected by people transitions instead [6, 33]. Intuitively, the commuting flows between regions also imply the flows of lifestyles. For example, it is reasonable to assume that consumers’ demands affect the shopping patterns in both the regions where they live and work. We assume that the stronger the interactions between two regions are, the more alike the lifestyles in both regions are. We define the interaction from a region to another region as the probability of a person moving from the former to the latter. We first collect two large human transition datasets in Beijing, and then apply a gravity model [26] to quantify the interactions. We incorporate the interactions between regions as an regularization term in our overall model. In comparison with several baseline methods, our model improves the accuracy by 10.8% and 12.6% in terms of the Mean Absolute Error (MAE) with 80% and 90% training data, respectively.

Our main contributions are summarized as follows:

- *Dealing with data sparsity:* We incorporate mobility patterns of regions as complementary information to address the sparsity issue of our shopping data. We leverage the underlying connection between mobility and shopping patterns, and use “where people go” to help infer “what people buy”. Moreover, to model the relations between regions and further make up for the spare data, we consider the interactions between regions in our overall model. A hybrid model is proposed to utilize several data sources or views. The experimental results validate that our model outperforms the baseline methods according to two standard metrics.

- *Multi-view Lifestyles:* We study the connection between different views of lifestyle particularly in terms of shopping and mobility patterns. The reasonable mappings between these views guarantee the effectiveness of incorporating mobility patterns to predict the shopping patterns. Moreover, interesting mappings between two views provide a better and novel understanding of human lifestyles in the city.
- *City-wide shopping patterns modeling:* We extract representative shopping patterns of regions in urban areas, and predict the level of each shopping pattern. Our results show the dramatic divergence in shopping demands over the city. This work sheds light on a better understanding of the urban infrastructure, and provides valuable information for various business opportunities.

2. RELATED WORK

2.1 Urban Region Divergences

Previous works on urban land use suggest various types of city regions [27, 3]. Singh et. al report their findings on noise levels of three types of city regions: residential, industrial, and commercial [22]. As social media has been being widely adopted in people’s daily lives, social media data becomes an important data source for related research. For example, geolocated tweets are used in [10] to characterize urban landscapes. In this work, the authors use tweet contents to identify business, leisure, nightlife, and residential regions in New York City.

Since urban data is usually noisy and sparse, researchers have to combine several data sources to dig meaningful knowledge. Yuan et. al [29] discover functional zones in Beijing. In this work, the authors cluster regions into several functional groups based on their transition patterns. POI information in a region is used as meta data for the region in this work. In [32], Yu et. al report their work on predicting the noise levels of New York City. City road structures, POI distribution, and check-in data are used in this work. Interesting correlations are found between multiple data sources and noise level. For example, universities in a region are quite responsible for “party & loud speak” noise. Similarly, the authors of [20] introduce road network and POI locations to infer the traffic volume in a region. [31] gives a nice survey on cross-domain data fusion in urban computing.

2.2 Human Lifestyles

Human lifestyles are well studied in sociology [2, 11]. In recent years, researchers have successfully utilized social media in research ventures related to lifestyle analysis [18, 1]. For example, Noulas et al. of [15] use Foursquare data to discover the behavioral habits of residents in London. Based on tweet contents, Sadiek et al. build a language model to detect the health of individuals [17].

Much effort has been made to linking multiple types of lifestyles. Cranshaw et. al [7] report that human mobility patterns are strongly connected to their social networks. In [34], the authors use check-in data to infer users’ demographic features. Furthermore, Yuan et. al linked several aspects of human lives, such as occupation, education level, and mobility patterns by studying the data from several social media platforms [30]. Inspired by these works, our key idea in this work is using information of one view (mobility) of lifestyles of a region to infer the other one (shopping).

2.3 Matrix Factorization Techniques

Matrix Factorization (MF) techniques, also known as Matrix Decomposition, are widely applied in recommendation systems to pre-

dict missing values in matrices [19]. A basic MF model factorizes the target matrix into a latent pattern matrix and a coefficient matrix, and uses the product of them as the estimation to missing values.

However, this approach usually suffers from data sparsity problems [12] – the information contained in the original matrix is not enough for a precise factorization. Our dataset has such a problem – a majority of rows in our target matrix are empty. Many techniques have been developed to deal with this problem [16, 35], including Collective Matrix Factorization (CMF) [21]. In a CMF model, supplementary information is provided in terms of matrices. Instead of decomposing only the target matrix, several matrices are decomposed together. By doing so, supplementary information helps to modify the factorization results, and produce a better prediction result. In this work, we apply a CMF model to combine shopping and mobility patterns. Social regularization is reported in [14]. In this work, Ma et. al consider the social connections between users in a recommendation system, and introduce the connections into a transitional MF model using a social regularization term. We introduce region interactions into our model using an interaction regularization term. The interactions are calculated using a gravity model [26]. This model assumes the human mobility in a city follows Newton’s law of gravity, and is widely applied in urban geography. Wang et. al [25] employ this model to calculate the influence between regions in their POI recommendation system.

3. OVERVIEW

Before we introducing our model, we first go through the workflow of this study in Figure 2. There are two main parts to this work: 1) shopping and mobility patterns extraction, and 2) predicting shopping patterns using a collective matrix factorization model with interaction regularization.

The first part starts from the raw browsing data and raw check-in data. We extract n shopping patterns (matrix P_s) and m mobility patterns (matrix P_m) from these two datasets. n and m are input parameters. From this process, we also obtain the weights on shopping patterns of different locations, as well as the weights on mobility patterns of LBSN users in the form of two coefficient matrices.

The second part starts from aggregating the locations and users into regions. After the aggregation, we obtain a region shopping pattern matrix R_s , a r by n matrix, and a mobility pattern matrix R_m , a r by m matrix. Here, r denotes the number of regions in the city. Also, we prepare the interaction matrix Q , a r by r matrix, which contains the interaction information between regions. We then apply collective matrix factorization with interaction regularization on R_s , R_m , and Q . This is based on the assumption that P_s can be approximately factorized to $R_l \times V_1^T$, meanwhile, P_m to $R_l \times V_2^T$. Three matrices are generated from this process: 1) R_l , a r by l region latent lifestyle matrix, which indicates the weights on latent lifestyles of regions. l is an input parameter denoting the number of latent lifestyles. 2) V_1^T , a l by n matrix. This matrix is the shopping view of latent lifestyles. 3) V_2^T , a l by m matrix, contains the mobility view of latent lifestyles. We predict the missing values in R_s with the corresponding elements of the product of R_l and V_1^T .

The basic idea behind collective factorization here is assuming that the weights on shopping patterns R_s and mobility patterns R_m of regions are both generated by l latent lifestyles. Each row in V_1^T explains a latent lifestyle with weights on shopping patterns, while a row in V_2^T explains a latent lifestyle with weights on mobility patterns. This process also reveals exquisite mappings between human shopping behaviors and movements. There are l columns in

R_l corresponding to l latent lifestyles. The i th column of R_l , for example, denotes the weights of regions on the i th row of V_1^T when using these two matrices to recover R_s . Meanwhile, i th column of R_l also denotes the weights of the regions on the i th row of V_2^T when using these two matrices to recover R_m . This intuitively implies the mapping between i th row in V_1^T and V_2^T – they are the two views of the same latent lifestyle.

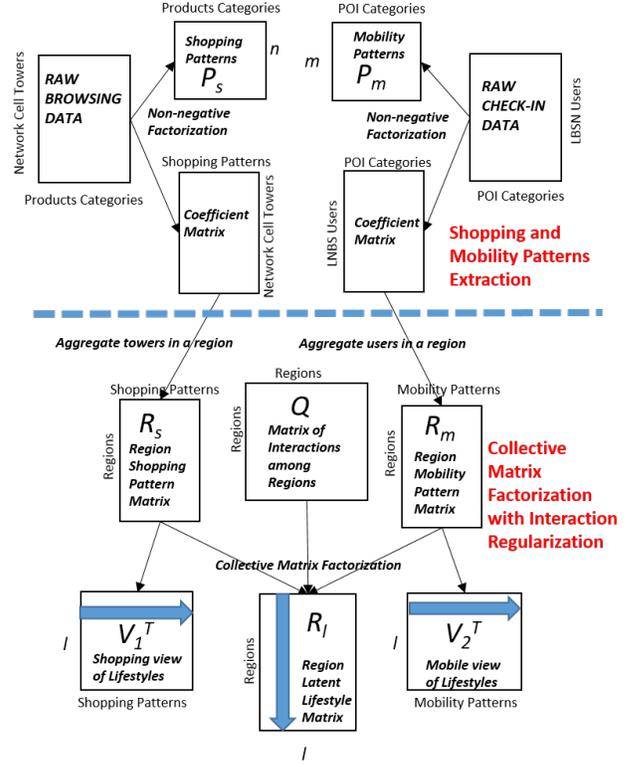


Figure 2: Our framework

4. SHOPPING AND MOBILITY PATTERNS OF REGIONS

We define a shopping pattern as a group of related product categories. For example, Furniture, Home Textiles, Large Home Appliances, and so on are extracted as a shopping pattern in this work. A mobility pattern is a group of related POI categories, such as a combination of School Dormitories, School Libraries, and so on. Shopping and mobility patterns are basic bricks of two views of lifestyles. In this section we explain how we extract these patterns.

4.1 Shopping Patterns Extraction

We collect the browsing log of shopping websites from the city of Beijing. Each log record is associated with a product category and a location identifier. There are no user identifiers in the data. Moreover, the location identifiers are not the precise locations where consumers visit product webpages, but the locations of the network cell towers their devices connect to.

Since we cannot identify users in our data, as a compromise, we use the data from a network cell tower as the basic unit for shopping pattern mining. A cell tower covers a relatively small area, hence, we assume that the people connected to a tower have similar shopping patterns. Note that, a city region could contain multiple or no network cell towers. We first aggregate the browsing

records from the same locations into a vector, we name this vector the browsing vector of a location. The length of a browsing vector is the number of product categories, and the value of a certain element is the browsing amount to that product category from a location. The browsing vectors are accumulated to form a browsing matrix. A row in this matrix indicates a location (a network cell tower), and a column indicates a product category. We then apply a Nonnegative Matrix Factorization (NMF) model on this browsing matrix, since NMF leads to interpretable results [13]. We obtain two result matrices through the processes: a shopping pattern matrix $P_s \in \mathbb{R}^{n \times c_s}$ and a coefficient matrix, where n is the number of shopping patterns, and c_s is the number of product categories. A row of P_s indicates the weights on product categories of a shopping pattern. The coefficient matrix indicates the weights on each shopping pattern of locations.

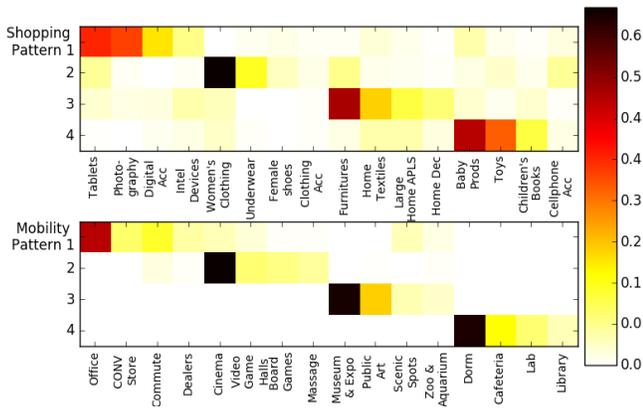


Figure 3: The top weighed categories of four shopping patterns (upper), and four mobility patterns (lower). Horizontal axes indicate product categories and POI categories, respectively. “Accs”, “Conv”, “Dorm”, “APLS”, “Intel”, “Prods”, and “Dec”, are short for “Accessories”, “Covenient”, “Dormiatroies”, “Applicances”, “Intellegent”, “Products”, and “Decoration”, respectively.

In Figure 3 we plot the top weighted product categories of several shopping patterns (upper part). Please note that the amount of categories of products and POIs is more than what we display in the figure. We can see that each shopping pattern has a specific emphasis on a group of related products, which naturally represents a class of demands of people. The first shopping pattern mainly focuses on digital products. Tablets, Photography Products, and Digital Accessories are the top three most weighted categories. The second pattern is dominated by Women’s Clothing, and other categories such as Underwear and Female Shoes have certain shares in the pattern. The third pattern shown in the figure is mainly about home arrangements and decorations, Furniture, Home Textiles, and Large Home Appliances are the most popular product categories in this pattern. The last shopping pattern has highest weights on Baby Products, Toys, and Children’s Books.

We then sum up the weights on shopping patterns of all the locations in a region to obtain the weights on the shopping patterns of a region. The result is represented by a matrix $R_s \in \mathbb{R}^{r \times n}$, where r is the number of regions in the city. The rows of this matrix indicate regions, and columns represent shopping patterns. Since many regions do not contain records at all, R_s is a sparse matrix. Our goal in this work is to fill in the empty elements of R_s .

4.2 Mobility Patterns Extraction

To extract human mobility patterns, we collect check-in data

from Jiebang², a popular LBNS in China. Each check-in is associated with a user ID, a POI category, and a latitude and longitude pair. We take a user as the basic unit for mobility pattern mining. The process is similar to shopping pattern extraction. We first aggregate users’ check-ins into vectors, named mobility vectors. Mobility vectors record users’ amounts of check-ins of all POI categories. We then accumulate mobility vectors into a mobility matrix, and apply NMF on it to extract a mobility pattern matrix $P_m \in \mathbb{R}^{m \times c_m}$, and a coefficient matrix. Here, m denotes the amount of mobility patterns, and c_m is the number of POI categories. P_m describes the ways in which people live from a view of spatial activities, and the coefficient matrix records the weights of users on these mobility patterns.

Figure 3 (lower part) shows four mobility patterns with their most weighted POI categories. Similar to shopping patterns, a mobility pattern also focuses on a specific group of categories. The first pattern indicates a mobility pattern centered at Offices, followed by Convenient Stores and Commute. The second mobility style is dominated by Cinemas. The next two categories are Video Game Halls and Board Games. This is clearly an entertainment related mobility pattern. The third one is more about outdoor activities, which assigns higher weights to Museum & Exhibition. The last mobility pattern mainly focuses on POIs related to schools or universities. Dormitories, Laboratories, and Libraries have high weights in this pattern.

Since LBNS users may leave footprints in several regions, it makes more sense to allocate the same proportion of user mobility patterns to the regions they checked in with the percentages of activities they spent in these regions. For example, if a user spends 40% check-ins in a region, then this user contributes 40% of her mobility patterns to this region. Therefore, we define the mobility patterns of a region as the weighted sum of all the users’ mobility patterns, who checked in this regions before. Matrix $R_m \in \mathbb{R}^{r \times m}$ is then formed, where the element on i th row and j th column denotes i th region’s weight on j th mobility pattern. Formally,

$$R_{m_{i,j}} = \sum_k w_{k,i} \times u_{k,j}, \quad (1)$$

where $w_{k,i}$ is the percentage of activity which user k spends in region i , and $u_{k,j}$ is this user’s weight on j th mobility pattern.

5. HYBRID MODEL FOR SHOPPING PATTERN PREDICTION

R_s contains the shopping pattern information of regions, and is a sparse matrix. Our objective is to fill the empties so as to predict shopping patterns of missing regions. In this section, we apply a Collective Matrix Factorization model with an interaction regularization term to achieve the goal.

5.1 Collective Matrix Factorization

We treat shopping and mobility patterns in a region as two different views generated by same lifestyles. In other words, we assume R_s and R_m are both generated from a matrix containing the latent lifestyle information of each region. Formally,

$$R_s \approx R_l V_1^T \quad (2)$$

$$R_m \approx R_l V_2^T \quad (3)$$

where $R_l \in \mathbb{R}^{r \times l}$ denotes the weights on l latent lifestyles of each region. $V_1^T \in \mathbb{R}^{l \times n}$ explains l latent lifestyles with n shopping

²<http://www.jiebang.com/>

patterns. Similarly, $V_2^\top \in \mathbb{R}^{l \times m}$ explains l latent lifestyles with m shopping patterns. Guided by the extra mobility information of R_m , we obtain more accurate R_l and V_1 . The prediction of missing values in R_s is achieved by multiplying R_l and V_1 . Following this idea, we apply a collective matrix factorization model to decompose R_s and R_m together. The objective function is formulated as:

$$L(R_l, V_1, V_2) = \frac{1}{2} \|I \circ (R_s - R_l V_1^\top)\|_F^2 + \frac{\lambda_1}{2} \|R_m - R_l V_2^\top\|_F^2 + \frac{\lambda_2}{2} (\|R_l\|_F^2 + \|V_1\|_F^2 + \|V_2\|_F^2), \quad (4)$$

where operator “ \circ ” means the entry-wise product, $\|\cdot\|_F^2$ denotes the Frobenius norm, and I is an indicator matrix with its entry $I_{ij} = 0$ if $R_{s_{ij}}$ is missing. The last form is for preventing overfitting. λ_1 and λ_2 are parameters controlling the contribution of each part.

An interesting knowledge we can learn from this approach is the mapping between two views. The i th row of V_1^\top and V_2^\top are both associated with the i th column in R_l , which implies that they represent the same latent lifestyle. However, only associating two views of lifestyles ignores the interactions between regions. We quantify the interactions using a gravity model, and introduce this factor in the objective function as a regularization term.

5.2 City-wide Interaction Regularization

5.2.1 Gravity Model

We define the interaction from region i to region j as the probability of a person moving from i to j . To compute this probability, we employ and adapt a gravity model [26], which is widely adopted in mobility analytics for a large population.

The gravity model states that the commuting flows from region i to j , denoted as $q_{i,j}$, are determined by 1) O_i , the number of individuals leaving region i , 2) D_j , the number of individuals arriving at region j , and 3) the distance between two regions, through a gravity-like law:

$$q_{i,j} = c \frac{(O_i)^a (D_j)^b}{\exp(g \cdot dis_{i,j})}, \quad (5)$$

where c is a constant, a and b are coefficients of O_i and D_j , respectively. g is a parameter that tunes the decay by distance. When $i = j$, this model computes the interactions within a region.

We fit this model using observed transition data to estimate the coefficients a , b , and g . Applying a logarithmic transformation to both side of (5), we obtain the following expression:

$$\ln q_{i,j} = a \ln O_i + b \ln D_j - g \cdot dis_{i,j} + \ln c \quad (6)$$

A multivariate regression method [5] is then employed on (6) to estimate the coefficients based on the observed data. We consider two types of commuting in a city: by taxi and by bus. Based them, we compute $q_{i,j}^B$ indicating the interaction by bus, and $q_{i,j}^A$ indicating the interaction by taxi. Figure 4 shows the interaction between a region (point X on the map) and other regions in the city. We then fill the quantified interactions into matrices Q^B and Q^A , where the element in the i th row and j th column is the spatial interaction from region i to region j by bus and taxi, respectively.

5.2.2 City-wide Interaction Regularization

Inspired by the social regularization reported in [14], we formulate the transitions between regions as a regularization term in our

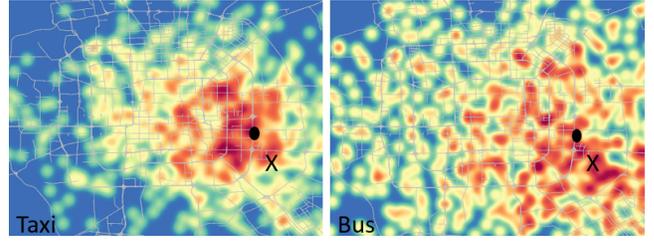


Figure 4: The heatmap of the interactions between region X with other regions in the city. A darker colored region indicates a higher probability of people in this region moving to region X.

objective functions. As mentioned in the introduction, interactions between regions transit lifestyles too. The more interactions between two regions, the more alike their lifestyles are. Based on this intuition, we propose the interaction regularization term as:

$$\frac{\alpha}{2} \sum_{i=1}^r \|R_{l_i} - \frac{1}{2} \sum_{* \in A, B} \frac{\sum_{j=1}^r Q^*(j, i) \times R_{l_j}}{\sum_{j=1}^r Q^*(j, i)}\|_F^2, \quad (7)$$

where $Q^*(j, i)$ denotes the interaction from region j to i by bus or taxi, and R_{l_i} indicates the lifestyles of i th region. We take two traffic types equally in the regularization, and average two types of interactions as the final interaction. α is a parameter that indicates the importance of this regularization term. This term limits the difference between regions’ lifestyles according to their interactions. In other words, a large $Q(j, i)$ leads to a strong influence from region j to region i . As a result of introducing this term, the cross-region influence is incorporate to our model.

5.3 Hybrid Model

We combine the collective matrix factorization model and interaction regularization term to obtain our final objective function:

$$L(R_l, V_1, V_2) = \frac{1}{2} \|I \circ (R_s - R_l V_1^\top)\|_F^2 + \frac{\lambda_1}{2} \|R_m - R_l V_2^\top\|_F^2 + \frac{\alpha}{2} \sum_{i=1}^r \|R_{l_i} - \frac{1}{2} \sum_{* \in A, B} \frac{\sum_{j=1}^r Q(j, i) \times R_{l_j}}{\sum_{j=1}^r Q(j, i)}\|_F^2 + \frac{\lambda_2}{2} (\|R_l\|_F^2 + \|V_1\|_F^2 + \|V_2\|_F^2), \quad (8)$$

There are generally no closed-form solutions for minimizing the above objective function. Therefore, we search for a local minimum by applying gradient descent in the lifestyle vector of each region R_{l_i} and matrix V_1, V_2 as follows

$$\begin{aligned} \frac{\partial L}{\partial R_{l_i}} &= \sum_{j=1}^r I_{i,j} (R_{l_i} V_{1j}^\top - R_{s_{i,j}}) V_{1j}^\top + \lambda_2 R_{l_i} \\ &\quad + \alpha \sum_{j=1}^r (R_{l_i} - \frac{1}{2} \sum_{* \in A, B} \frac{\sum_{j=1}^r Q^*(j, i) \times R_{l_j}}{\sum_{j=1}^r Q^*(j, i)}) \\ \frac{\partial L}{\partial V_1} &= [I \circ (R_l V_1^\top - R_s)]^\top R_l + \lambda_2 V_1 \\ \frac{\partial L}{\partial V_2} &= \lambda_1 (R_l V_2^\top - R_m)^\top R_l + \lambda_2 V_2. \end{aligned} \quad (9)$$

By optimizing the objective function, we obtain R_l, V_1 , and V_2 . According to Equ. 2, we predict the missing values in R_s with the

product of R_s and V_1^\top . Also, the corresponding rows of V_1^\top and V_1^\top explain same lifestyles from two views. We summarize our algorithm in Algorithm 1.

Algorithm 1: Optimization of the Hybrid Model

Input: region shopping pattern matrix R_s , region mobility pattern matrix R_m , region interaction matrix Q^A and Q^B , and parameters: λ_1 , λ_2 , and α

Output: region lifestyle pattern matrix R_l , lifestyle-shopping matrix V_1 , and lifestyle-mobility matrix V_2

- 1 initialize R_l , V_1 , and V_2 ;
- 2 //T is maximum iterations, ϵ is a stopping threshold
- 3 **while** $t < T$ and $L_t - L_{t+1} > \epsilon$ **do**
- 4 calculate the objective function L_t based on Equ. 8;
- 5 calculate $\frac{\partial L}{\partial R_{l_i}}$, $\frac{\partial L}{\partial V_1}$, and $\frac{\partial L}{\partial V_2}$ based on Equ. 9;
- 6 $\gamma = 1$;
- 7 //search for the maximal learning rate;
- 8 **while**
 $L(R_{l_i} - \gamma \frac{\partial L}{\partial R_{l_i}}, V_1 - \gamma \frac{\partial L}{\partial V_1}, V_2 - \gamma \frac{\partial L}{\partial V_2}) \geq L(R_l, V_1, V_2)$
do
- 9 $\gamma = \gamma/2$;
- 10 update R_{l_i} , V_1 , and V_2 based on Equ. 9;
- 11 $t = t + 1$;
- 12 **return** R_l , V_1 , and V_2

6. EVALUATION

6.1 Data

We part the city into 1km by 1km grids aligned with the latitude and longitude lines. We take the urban area within Beijing 5th Ring Road, and there are $29 \times 30 = 870$ grids in total, and we do not have any browsing records for 548 (62.9%) of them. Each grid is taken as a region. User online browsing data is used for extracting R_s , the shopping patterns of each region. Check-in data from a popular LBSN is used to extract R_m , the mobility patterns of regions. To learn the interactions between regions, we utilize two large datasets of taxi and bus transitions in the city. We describe these datasets as follows:

- *online browsing dataset:* We collect the online shopping browsing logs of Internet Explorer (IE) users, who have granted permission for recording their log data. Our dataset is collected from Nov to Dec 2015, and contains all the log records generated in Beijing. Each record is associated with a product category and a location identifier. There are more than 250 product categories in total. Many constraints are set up in the collection process to protect user privacy, which make this study more challenging. First, there are no user identities in the data. As a result, we are not able to trace any personal browsing history. Also, the locations of records are not precise coordinates, but the locations of network cell towers. This constraint prevents us from locating where the browsing behaviors are taking place exactly.
- *check-in dataset:* We first collect more than 2 million check-in data in Beijing from Aug 2011 to Jan 2013. A check-in contains a time-stamp, a location, and a user ID. Each location is associated with a POI category, and there are around 200 categories in total. Since tourists generate a lot of check-ins in a large city like Beijing, we cannot use these check-ins

directly. Following the user IDs, we download the profiles of these users. A profile contains basic information about a user including their home city. We then collect the users, whose home city is Beijing, and aggregate their check-ins. After doing so, we obtain 1.5 million check-ins generated by 43 thousand users.

- *bus dataset:* This dataset contains 3 million bus-trip records. The data was collected in Beijing from August 2012 to May 2013, and each record is a 5-tuple consisting of a user ID, boarding and alighting locations, and alighting time.
- *taxi dataset:* This dataset consists of 1.9 million taxi trips in Beijing from March 2011 to August 2011. Each trip includes the location of the origin and destination, as well as the boarding and alighting time.

The check-in data and the transition data are not as up-to-date as our browsing data dataset. Since the urban structure of Beijing has not changed much in recent years, this allows us to apply these data as complementary information sources in our prediction.

6.2 Metrics

We apply two popular metrics to evaluate our model, the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). RMSE and MEA are defined as:

$$RMSE = \sqrt{\frac{1}{T} \sum_{i,j} (R_{s_{i,j}} - \hat{R}_{s_{i,j}})^2}, \quad (10)$$

$$MEA = \frac{1}{T} \sum_{i,j} |R_{s_{i,j}} - \hat{R}_{s_{i,j}}|, \quad (11)$$

where $R_{s_{i,j}}$ denotes the true value of j th shopping pattern in i th region, and $\hat{R}_{s_{i,j}}$ is the estimation of it. T is the number of tested values. Clearly, a lower RMSE or MEA value indicates a better performance.

6.3 Baselines

We compare our model with three baselines described as follows.

- *Matrix Factorization (MF):* this method decomposes R_s into R_l and V_1 directly, and uses the product of two results matrices as the prediction.
- *Collective Matrix Factorization (CMF):* in this baseline, we factorize R_s and R_m together to obtain R_l , V_1 , and V_2 , without using interaction information.
- *CMF with neighboring information:* since a region usually has the most interactions with the regions surrounded it, in this method, we assume that a region is only affected by neighboring regions. Accordingly, the interaction term is reformed as:

$$\frac{\alpha}{2} \sum_{i=1}^r \|R_{l_i} - \frac{\sum_{j \in neighbors} R_{l_j}}{\|neighbors\|}\|_F^2, \quad (12)$$

where $neighbors$ is the set of indexes of the regions that surround region i , and $\|neighbors\|$ denotes the amount of neighbors. We take the eight surrounding regions as the neighbors of a region. The comparison between our model and this baseline shows the advantage of introducing the interactions between all regions over simply considering local interactions with neighbors.

Training Size	80%		90%	
Metrics	RME	MEA	RME	MEA
MF	0.235	0.443	0.213	0.420
CMF	0.230	0.408	0.203	0.400
improve	2.12%	7.90%	4.69%	4.76%
CMF + N	0.226	0.398	0.200	0.374
improve	1.73%	2.5%	1.47%	6.5%
CMF + I	0.223	0.394	0.197	0.367
improve	1.30%	1.00%	1.50%	1.87%
Total	5.1%	10.81%	7.5%	12.61%

Table 1: RME and MEA values of each model. “CMF + N” denotes CMF model plus neighboring information, and “CMF + I” denotes CMF model plus interaction information. “improve” denotes the improvement of accuracy of each model compared with the model before it. “Total” indicates the improvement of accuracy using our model compared with MF model.

6.4 Results

In many previous works on matrix factorization [14], testing data are formed by randomly selected non-zero elements. These selected non-zero elements are compared with the prediction results to evaluate models. In our problem, R_s contains many empty rows. To test if our model is able to predict the regions that we do not have any information about, we set up a harsher testing strategy. We first randomly select non-zero rows instead of elements from the original matrix to form the testing data. We then compare the non-zero element in these rows with the prediction results. We let $n = 30$, $m = 40$, and $l = 10$, then train the models on 80% and 90% of the original data, respectively. We run the experiment ten times, and report the average results in Table 1.

The table shows that our model outperforms all baseline models, and brings a dramatic improvement to the basic matrix factorization model on both metrics. First, introducing mobility patterns into the task with a CMF model brings a large boost to MF. This indicates that the connection between human shopping patterns and mobility patterns works effectively for prediction. Our model learns the connection from the known regions, and leverages it to predict the regions where shopping patterns are unknown. Adding neighboring information to the model also increases performance. Because people prefer to move to a close place [23], a region usually affects the neighboring regions the most. Furthermore, replacing the neighboring information with interactions between regions brings another reasonable boost to accuracy. This proves our assumption that the interactions between regions also spread the shopping patterns of them. In the next section, we report our prediction results, and discuss the geographical distributions of different shopping patterns.

7. MULTIPLE VIEWS OF LIFESTYLES

In our results, V_1^\top and V_2^\top indicate shopping and mobility views of latent lifestyles. The reasonable mappings between these views guarantee the effectiveness of incorporating mobility patterns to predict the shopping patterns. In this section, we discuss the mappings discovered in the results. Note that, a shopping view of lifestyle consists of a group of related shopping patterns, and a shopping pattern is formed by a group of related product categories. It is the same for mobility views of lifestyle. Due to the length limit, we pick four latent lifestyles as examples. We list the top three shopping/mobility patterns for the latent lifestyles in Table 2. For each shopping/mobility pattern, we list the top three product/POI categories.

Lifestyle 1 describes the typical life of urban white-collars. The top weighted mobility patterns mainly focus on work (Office, Convenient Stores, and Commute), home (Apartment & Communities, Commute), and shopping places (Shopping Malls, Cosmetics Stores, and Sporting Stores). Interestingly, instead of office appliances, the most popular shopping pattern is about women’s clothing, underwear, and women shoes. This may indicate the online shopping behaviors of office clerks while working. The second shopping pattern consists of Laptops, Office Appliances & Stationery, and Tablets, which are usually needed for work. The mapping between these two views of this lifestyle explains the overlapping of the two geographical distributions shown in Figure 1. The demands for office appliances and the check-ins at offices can be explained as generated by this latent lifestyle. Therefore, the high level of this lifestyle in a region naturally results in the high levels of both patterns.

Lifestyle 2 shows a strong emphasis on digital and electric products from its shopping view. The top shopping pattern of this lifestyle contains Tablets, Photography, and Digital Accessories, and the third shopping pattern contains Peripheral and Network Products. We can see from its view of mobility that, a group of POI categories lead by Digital Appliance Stores is one of the top mobility patterns. Also, this lifestyle has a high demand for books – the second shopping pattern is dominated by Books. Correspondingly, we observe that the most popular mobility pattern in this lifestyle contains Public Libraries, Book Stores, and Public Arts. It also shows a preference for gyms (Gym, Stadiums, and Billiard Parlor) of people of this lifestyle.

Lifestyle 3 is quite life-oriented. We observe the shopping pattern of Furniture, Home Textiles, and Large Home Appliances is listed in the second place in this lifestyle. This pattern implies the need for home arrangements and decorations, and it matches the first mobility pattern of the mobility view, which is lead by Furniture Stores. Also, auto accessories are of high demand in this lifestyle. This corresponds to the visits to garages and dealers in the mobility view. Moreover, the people of this lifestyle have a relatively high preference for traveling, because they pay more visits to scenic spots and public squares (second mobility pattern). This mobility pattern, conversely, implies the demands of auto accessories of the shopping view. The top shopping pattern consists of computer accessories, cellphone accessories, and other computer products, which decries a general demand for electric products.

Lifestyle 4 has the mobility patterns that mainly focus on entertainment (Cinemas, Video Game Halls, and Board Games), shopping places (Supermarkets, Sporting Stores, and Convenient Stores), and schools or universities (School Dormitories, School Laboratories, and School Libraries). From the view of mobility, we suppose that many students could have high rates on this lifestyle. As to its shopping view, Men’s Clothing, Male Shoes, and Clothing Accessories form the top shopping pattern of this lifestyle, followed by digital products, and then books. Women’s clothing is also quite popular in this lifestyle, but not listed in the top three. For students, it is natural to have high demand for books. This is also reflected in our prediction results – the level of the shopping pattern of book is high in educational regions (see Section 8.3). Moreover, teenagers usually are more interested in digital products, which explains the digital products shopping pattern.

The reasonable mapping between two views explains why our model works effectively: a region has certain latent lifestyles, and the views of lifestyles are actually connected. Therefore, we are able to utilize the data of another view (mobility) to prediction our target (shopping).

Lifestyle	Lifestyle 1		Lifestyle 2	
Views	Shopping View	Mobility View	Shopping View	Mobility View
First Pattern	Women Clothing Underwear Women Shoes	Offices Convenient Stores Commute	Tablets Photography Products Digital Accessories	Public Libraries Book Stores Public Art
Second Pattern	Laptops Office APLS & Stationery Tablets	Apartments & Communities Commute Other POIs	Books Children’s Books Health & Fitness	Gymnasium & stadiums Billiard Parlors School Gyms
Third Pattern	Cellphones Cellphone Plans Cellphone Accessories	Shopping Malls Cosmetics Stores Sporting Stores	Peripheral Products Network Products Underwear	Digital Appliance Stores Commute Sporting Stores
Lifestyle	Lifestyle 3		Lifestyle 4	
Views	Shopping View	Mobility View	Shopping View	Mobility View
First Pattern	Computer Accessories Cellphone Accessories Other Computer Products	Furniture Stores Cosmetics Stores Garages & Dealers	Men Clothing Men Shoes Clothing Accessories	Cinemas Video Game Halls Board Games
Second Pattern	Furniture Home Textiles Large Home Appliances	Public Squares & Blocks Scenic Spots Other POIs	Tablets Photography Products Digital Accessories	Supermarkets Sporting Stores Convenient Stores
Third Pattern	Auto Accessories Multimedia products Cellphone Accessories	Cinema Video Game Halls Board Games	Books Children’s Books Health & Fitness	School Dormitories School Laboratories School Libraries

Table 2: Four lifestyles from both shopping and mobility views. For each view we list the top four most weighted shopping/mobility patterns, and for each pattern we list the top four most weighted product/POI categories. “APLS” is short for “Appliances”.

8. PREDICTED SHOPPING PATTERNS IN BEIJING

The geographical distributions of shopping patterns not only tell the various demands of regions, but also uncover the urban structure of a city. For example, a high level of shopping behaviors implies a high level of commerce. Furthermore, the level of a specific shopping pattern could reveal key features of regions. In this section, we display our prediction results on a map, and compare the levels of these shopping patterns city-wide. Clear differences between the distributions reveal interesting insights to the city.

8.1 Business Centers of Beijing

We plot the prediction results of two shopping patterns on a Beijing subway map in Figure 5. The shopping pattern on the left is mainly about office appliances, and the one on the right is mainly about women’s clothing. The distributions of these two patterns are quite similar, except that the pattern of women clothing is more widely spread. We suppose that this implies the online shopping behaviors of office clerks during working hours. As we showed in last section, office appliances and women’s clothing are clustered into the shopping view of the same latent lifestyle in our results (see Table 2, shopping view of lifestyle 1).

More importantly, the distribution of office-related shopping patterns implies the locations of business centers in the city. Several regions have much higher levels on this shopping pattern. Two groups of regions can be observed on the map. The regions in the first group are located in the middle of the city, and they line up horizontally. The line linking these regions aligns with Beijing Subway Line 1, which is highlighted on the figures with a dash line. Line 1 is not only the oldest, but also the most heavily used line of the Beijing Subway³. It links the earliest and most important economic centers in the city, such as Guomao, Xidan, and Wangfujing. Moreover, the locations of the highest levels of shopping patterns



Figure 5: The geographical distributions of two shopping patterns on a Beijing subway map. The top three weighted product categories of them are: Laptops, Office Appliances & Stationery, and Tablets (left); Women Clothing, Underwear, and Women Shoes (right).

on the line match with the subway stations of Line 1. This concurs with the conclusions on the positive effects of subway stations on the economies of their surroundings [8]. Also, it can be observed that, the line is split by Tian’anmen Square and the Forbidden City, two landmarks in Beijing. The levels of this pattern in the east part on this line are higher than the west. This agrees with the fact that the eastern areas are more developed than the western areas in Beijing. According to the statistics from the Beijing Municipal Bureau of Statistics⁴, this eastern district (Chaoyang district) has the highest Gross Domestic Product (GDP) among all districts in Beijing. The second group consists of the regions in the northwestern corner of the city – Haidian district (circled on the map). This is a high-tech district hosting many software and computer-technology companies, as well as several higher education institutes⁵. Our results denote that this district has grown to another business center of the city. This agrees with the fact that Haidian takes the second place of the GDP ranking of Beijing.

⁴<http://old.bjstats.gov.cn/esite/>

⁵https://en.wikipedia.org/wiki/Haidian_District

³https://en.wikipedia.org/wiki/Beijing_Subway

By comparing our results with real world locations, we show that our prediction successfully locates the business centers in an urban area.

8.2 Central VS. Peripheral Regions

Another two shopping patterns are visualized in Figure 6. The one on the left is about house arrangements and decorations, since Furniture, Home Textiles, and Large Home Appliances are the most weighted product categories of this pattern. The other shopping pattern on the right weights Baby Products, Toys, and Children’s Books the most. Not surprisingly, the business centers that we discussed in last section have high levels of these patterns. Because many people work in these areas, the high population base of potential online shopping consumers naturally results in a high level of demand.



Figure 6: The geographical distributions of two shopping patterns. The top three weighted product categories of them are: Furniture, Home Textiles, and Large Home Appliances (left); Baby Products, Toys, and Children Books (right).

Compared with the shopping patterns of office appliances and women’s clothing (See Figure 5), these two patterns are more widely spread. The levels of the pattern of home appliances are relatively high in the peripheral areas of Beijing. As for the patterns of baby products, there are two observations. First, demand for it covers the entire city, which implies that this is a general demand in daily life. Second, the levels in the peripheral regions of this pattern also appear higher than the central areas (except for the business centers). The difference between central and peripheral areas reflected by our results can be explained by the process of urban development of Beijing. The central areas of Beijing are traditional districts. These districts are well developed, and the average age of the residents in these areas are relatively high. While the peripheral regions are still under developing. Younger people tend to live in these areas. Also, many ongoing real estate projects are located there. Since products for house arrangements and decorations are usually for new houses, it is reasonable to observe high levels in peripherals. Meanwhile, the baby products are usually purchased by young parents, and the people living in peripheral regions are relatively young, which explains the higher level of the baby product pattern in these regions.

8.3 Unique Shopping Patterns of Regions

It is not always true that business centers have the highest levels of all shopping patterns. We observe some shopping patterns that are only predicted popular in certain regions. The left plot of Figure 7 describes the geographical distribution of the group for Desktops, Computer Accessories, and Network Accessories. Although this shopping pattern is quite work-related, Haidian district has a much higher level than other business centers. This quite makes sense, because this district is the most developed IT area in Beijing (even in China). The products included in this shopping pattern are especially needed for the IT sector.

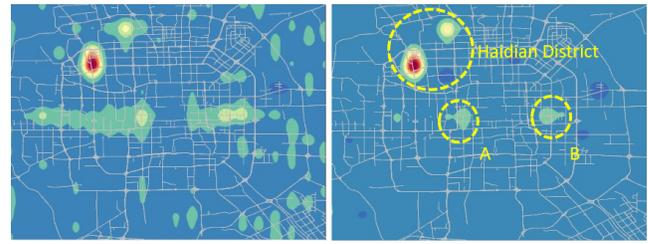


Figure 7: The geographical distributions of two shopping patterns. The top three weighted product categories of them are: Desktops, Computer Accessories, and Network Accessories (left); Books, Children’s Books, and Health & Fitness (right).

Since the need for computers are universal in modern jobs, we can still observe other business centers in the middle of the city in the figure. While another shopping pattern on the right of Figure 7 has the narrowest geographical distribution among all of them. There are only a few regions have relatively high levels for this shopping pattern. Book is the dominant category in this pattern. Due to the natural demands for books of students and technicians, Haidian district has the highest demands for books comparing with other regions. Region A (circled) has the largest book store in Beijing. We suppose this is because people would search books online for extra information such reviews when visiting a bookstore. Region B (circled) is the Central Business District (CBD) of Beijing. It appears that this region also has a relatively high demands for books.

9. CONCLUSION AND FUTURE WORK

In this paper, we focus in predicting shopping patterns of different city regions using sparse data. We incorporate two types of extra information to help solve such a challenging problem. First, we connect the shopping patterns with the mobility patterns in a region, in other words, use “where people go” to help infer “what people buy”. Second, we model the interactions between regions, and leverage the information of known regions to infer the shopping patterns in unknown regions. A hybrid model is introduced for the prediction task. The experiment results indicate that incorporating extra information in the task leads to a significant boost (5%+ in terms of RME, and 10%+ in terms of MEA) of accuracy for our prediction task. More importantly, our prediction results reveal key functional characteristics of urban areas, including the locations of business centers, the difference between central and peripheral areas, and the unique demands of certain city regions.

There are mainly two directions for our future work. First, we would like to introduce the temporal dimension into the shopping pattern prediction task. Since people’s demands vary along time, it is more valuable to predict “what people buy at when”. Second, we would like to introduce more views of lifestyles, such as personalities, occupations, and so on, to our framework. There are many views to observe how people live. It is therefore beneficial to link more views, and obtain a better and more comprehensive understanding of human lifestyles.

10. ACKNOWLEDGMENT

We would like to thank the Xerox Foundation. We also thank Xitong Yang for helping with the plots in this paper.

11. REFERENCES

- [1] S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through twitter. *arXiv preprint arXiv:1412.4361*, 2014.

- [2] H. L. Ansbacher. Life style: A historical and systematic review. *Journal of individual psychology*, 23(2):191, 1967.
- [3] J. Antikainen. The concept of functional urban area. *Informationen zur Raumentwicklung*, 7:447–456, 2005.
- [4] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record*, volume 26, pages 255–264. ACM, 1997.
- [5] A. J. Burnham, R. Viveros, and J. F. MacGregor. Frameworks for latent variable multivariate regression. *Journal of chemometrics*, 10(1):31–45, 1996.
- [6] S. Chawla, Y. Zheng, and J. Hu. Inferring the root cause in road traffic anomalies. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 141–150. IEEE, 2012.
- [7] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM UbiComp*, pages 119–128. ACM, 2010.
- [8] D. N. Dewees. The effect of a subway on residential property values in toronto. *Journal of Urban Economics*, 3(4):357–369, 1976.
- [9] R. Du Preez, E. Visser, and L. Zietsman. Profiling male apparel consumers: lifestyle, shopping orientation, patronage behaviour and shopping mall behaviour. *Management Dynamics*, 16(1):2, 2007.
- [10] V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez. Characterizing urban landscapes using geolocated tweets. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 239–248. IEEE, 2012.
- [11] T. Kamphorst. Leisure and lifestyle. *World Leisure & Recreation*, 32(2):31–32, 1990.
- [12] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.
- [13] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [14] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *Proceedings of the fourth ACM WSDM*, pages 287–296. ACM, 2011.
- [15] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM*, 11:70–573, 2011.
- [16] S.-T. Park and W. Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 21–28. ACM, 2009.
- [17] A. Sadilek, S. Brennan, H. Kautz, and V. Silenzio. nemesi: Which restaurants should you avoid today? In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.
- [18] A. Sadilek and H. Kautz. Modeling the impact of lifestyle on health at scale. In *Proceedings of the sixth ACM WSDM*, pages 637–646. ACM, 2013.
- [19] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th WWW*, pages 285–295. ACM, 2001.
- [20] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu. Inferring gas consumption and pollution emission of vehicles throughout a city. In *Proceedings of the 20th ACM SIGKDD*, pages 1027–1036. ACM, 2014.
- [21] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD*, pages 650–658. ACM, 2008.
- [22] B. B. Singh and V. Jain. A comparative study of noise levels in some residential, industrial and commercial areas of delhi. *Environmental monitoring and assessment*, 35(1):1–11, 1995.
- [23] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [24] P. Wang, J. Guo, Y. Lan, J. Xu, and X. Cheng. Your cart tells you: Inferring demographic attributes from purchase data. 2015.
- [25] Y. Wang, N. J. Yuan, D. Lian, L. Xu, X. Xie, E. Chen, and Y. Rui. Regularity and conformity: Location prediction using heterogeneous mobility data. In *Proceedings of the 21th ACM SIGKDD*, pages 1275–1284. ACM, 2015.
- [26] A. G. Wilson. A statistical theory of spatial distribution models. *Transportation research*, 1(3):253–269, 1967.
- [27] J. Xiao, Y. Shen, J. Ge, R. Tateishi, C. Tang, Y. Liang, and Z. Huang. Evaluating urban expansion and land use change in shijiazhuang, china, by using gis and remote sensing. *Landscape and urban planning*, 75(1):69–80, 2006.
- [28] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR*, pages 325–334. ACM, 2011.
- [29] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD*, pages 186–194. ACM, 2012.
- [30] N. J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu, and X. Xie. We know how you live: exploring the spectrum of urban lifestyles. In *Proceedings of the first ACM conference on Online social networks*, pages 3–14. ACM, 2013.
- [31] Y. Zheng. Methodologies for cross-domain data fusion: an overview. *Big Data, IEEE Transactions on*, 1(1):16–34, 2015.
- [32] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, and E. Chang. Diagnosing new york city’s noises with ubiquitous data. In *Proceedings of the 16th ACM UbiComp*, pages 715–725. ACM, 2014.
- [33] Y. Zheng, H. Zhang, and Y. Yu. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. *Submitted to ACM SIGSPATIAL*, 2015.
- [34] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the Eighth ACM WSDM*, pages 295–304. ACM, 2015.
- [35] K. Zhou, S.-H. Yang, and H. Zha. Functional matrix factorizations for cold-start recommendation. In *Proceedings of the 34th international ACM SIGIR*, pages 315–324. ACM, 2011.