Transfer Learning for Urban Computing: A Case Study for Optimal Retail Store Placement

Ningyu Zhang
Department of Computer Science
Zhejiang University Hangzhou China
zhangningyu@zju.edu.cn

Huajun Chen[*]
Department of Computer Science
Zhejiang University Hangzhou China
huajunsir@zju.edu.cn

Xi Chen
Department of Computer Science
Zhejiang University Hangzhou China
chenxi@zju.edu.cn

ABSTRACT

In recent years, big data analysis has been applied to the design and development of smart cities, which creates opportunities as well as challenges. It is necessary to retrieve a large amount of social media data and physical sensor data for this purpose. However, different cities have different infrastructures and populations, resulting in the sparsity of some types of data, such as social media data. In this paper, we propose a method for transfer learning between smart cities and apply it to optimal retail store placement owing to its importance in the success of a business. Traditional approaches to the problem have considered demographics, revenue, and aggregated human flow statistics from nearby or remote areas; however, the acquisition of relevant data is usually expensive. The rapid growth of location-based social networks in recent years has led to the availability of fine-grained data describing the mobility of users and popularity of places. However, circumstances vary from one city to another. Furthermore, the number of sensors may not be sufficient to cover all the relevant areas of a particular city. In such cases, it would be useful to transfer knowledge to small cities. We study the predictive power of various machine-learning features with regard to the popularity of retail stores in a city by using datasets collected from open data sources in several big cities. In addition, we use a multi-view discriminant transfer learning method to transfer knowledge to small cities. The results of experiments involving cities in China confirm the effectiveness of the proposed framework.

Keywords

Urban Computing; Optimal Retail Location; Smart City; Transfer Learning

1. INTRODUCTION

Urban computing, which aims to tackle urban problems by using city-generated data (e.g., traffic flow, human mobility, and geographical data), connects urban sensing, data management, data analytics, and service provision into a recurrent process to continuously and unobtrusively improve the quality of life, city operation systems, and the environment. For instance, the geographical placement of a retail store or new business has been of prime importance since the establishment of the first urban settlements in ancient times, and it assumes the same importance from the viewpoint of modern trading and commercial ecosystems in today's cities. A new coffee shop that is set up in a street corner may thrive with hundreds of customers, but it may close in a matter of months if it is set up a few hundred meters down the road. Nevertheless, infrastructure statistics are not sufficient for evaluating investment values. For example, the noise and pollution associated with train/bus systems can lower the value of a coffee shop. Thus, the utility of infrastructure statistics is limited. Moreover, such statistics are rarely dynamic and do not adequately reflect the changing profile of a city.

In contrast, from the perspective of urban computing, more dynamic and information-rich data can be accumulated with the development of mobile, internet, and sensor technologies. For example, people may post comments and ratings for places of interest (POIs; e.g., schools, restaurants, and shopping centers) via mobile apps after consumption. Moreover, mobility data such as smart card transactions and taxi GPS traces consist of both trajectories and consumption records of residents' daily commutes. If properly analyzed, these data (e.g., user reviews and location traces) can serve as a rich source of intelligence for determining optimal retail store placement.

Indeed, these retail-related dynamic data generated by users could better reflect values of placement than urban infrastructure statistics. In general, if people have good opinions of a store, the demand for this store as well as its investment value will be high. The challenge is how to uncover people's opinions of a store. In fact, the opinions of users can be mined from (1) online user reviews and (2) offline urban regional data. Specifically, online reviews (e.g., Dianping/Weibo ratings) contain explicit opinions regarding places surrounding a store. For example, the quality of a neighborhood can be partially approximated by the ratings of business venues, such as overall rating, service rating, and environment rating. On the other hand, offline urban regional data near a store not only encode the static statistics of

[*]Corresponding author

UrbComp'15, August 10, 2015, Sydney, Australia..

urban infrastructure but also reflect residents' implicit opinions of the neighborhood. For example, the arriving, transition, and departing volumes of taxies and buses indicate the mobility density of a neighborhood; the average velocity of taxies and buses indicates the degree of traffic congestion or accessibility; and the price of real estate and the traffic congestion index indicate whether the facility planning is balanced. All these indications provided by dynamic user-generated data reveal important facets of a store that are of great concern to customers and convey the implicit user opinions of a neighborhood. Therefore, we consider and mine both the explicit opinions from user reviews and the implicit opinions from urban regional data to enhance the evaluation of optimal retail store placement.

However, different cities have different infrastructures and populations, resulting in the sparsity of some types of data for smart cities[2]. For example, it is relatively easy to obtain heterogeneous data such as online user reviews in a metropolis because of its large population and infrastructure. However, small towns have small populations, and hence, relatively low social media activity. Therefore, it is difficult to assess optimal retail store placement on the basis of such data from small cities. On the other hand, large cities have been extensively modeled for numerous applications through big data analysis. In this paper, we propose a method to transfer knowledge between smart cities and apply it to optimal retail store placement.

Specifically, transfer learning aims to extract common knowledge across domains such that a model trained on one domain can be adapted effectively to other domains. In reality, different cities are equivalent to different domains, and online and offline data can be regarded as two different views (social view and physical view, respectively). Given a set of candidate areas in a city for opening a store, our aim is to identify the best ones in terms of their potential to attract a large number of users (i.e., to become popular). We formulate this problem as a rank problem, where, by extracting a set of features, we seek to exploit them to assess the retail quality of a geographic area. More specifically, we handle social and physical views through a multi-view discriminant transfer learning method. We adopt autoencoders to construct a feature mapping from an original instance to a hidden representation[26], and we use the source domain data to train a classifier for predictions on the target domain. Based on the framework, a particular solution is proposed to learn the hidden representation and classifier simultaneously.

The major contributions of this paper are as follows:

1) We propose a multi-view discriminant transfer learning method for urban computing between smart cities. We apply this method to optimal retail store placement in order to transfer knowledge from large cities to small ones to improve accuracy. This addresses the issue of what to transfer and how.

2) We study the factors affecting transfer learning in order to select the transfer candidates. We analyze the changes in different cities, and we formulate rules to select transfer candidates. This addresses the issue of when to transfer knowledge.

3) We evaluate our approach using various data sources from the Web, including traffic data, bus data, user comments in China, in order to verify the effectiveness of the proposed framework.

The remainder of this paper is organized as follows. Section 2 briefly reviews existing studies on retail store placement and transfer learning. Section 3 presents preliminary definitions and the problem statement. Section 4 describes social and physical view analyses as well as the proposed multi-view discriminant transfer learning method. Section 5 presents the results of our experiments. Finally, Section 6 summarizes our findings and concludes the paper with a brief discussion on the scope for future work.

2. RELATED WORK

2.1 Urban Computing

The dynamics of a city (e.g., human mobility and the number of changes in a POI category) may indicate trends in the city's economy[24][23][13][14][15]. For instance, the number of movie theaters in Beijing kept increasing from 2008 to 2012. This could mean that an increasing number of Beijing's residents preferred to watch movies in movie theaters. In contrast, some categories of POIs may vanish in a city, signifying a downturn in business. Likewise, human mobility could indicate the unemployment rate in some major cities and therefore facilitate the prediction of stock market trends. Thus, human mobility combined with POIs can help determine the placement of some businesses.

Land economy community research has concentrated on spatial interaction models, which are based on the assumptions that (1) the intensity of interaction between two locations decreases with their distance and (2) the usability of a location increases with the intensity of use and the proximity of complementary arranged locations. However, it has been shown that the applicability of these models is limited to agglomerations such as large shopping centers, and their predictive accuracy decreases when smaller, specialized stores are considered.

Karamshuk et al.[12] studied the problem of optimal retail store placement in the context of location-based social networks. They collected and analyzed human mobility data from Foursquare to understand how the popularity of three retail store chains in New York is shaped in terms of the number of check-ins. A diverse set of data mining features were evaluated, modeling spatial and semantic information about places and patterns of user movements in the surrounding area. Thus, among these features, places of special interest to users (i.e., train stations or airports) and retail stores of the same type as the target chain (i.e., coffee shops or restaurants), which encode the local commercial competition in an area, are the strongest indicators of popularity. The problem arises when we cannot obtain sufficient online data in some small cities.

With respect to previous work in the general area, in this paper, we examine how the problem can be framed through transfer learning. The richness of information provided by the above-mentioned services in big cities could enable us to study the retail quality of an area in a fine-grained manner: various types of geographic, semantic, and mobility information can not only complement traditional techniques but also form the basis for a new generation of business analytics driven by online data.

2.2 Transfer Learning

Transfer learning is what happens when someone finds it much easier to, for example, learn to play chess having already learned to play checkers, or recognize tables having already learned to recognize chairs, or learn Spanish having already learned Italian[1][7].

Transfer learning models data from related but not identically distributed sources[17][20][19][18][3]. Multi-view learning has been studied extensively in single-domain settings, such as in co-training[6][22][11]. However, multi-view transfer has not attracted much attention. Chen et al.[4] proposed Co-training for Domain Adaptation (CODA), a pseudo multi-view algorithm with only one view for original data that may not be effective for a real multi-view case. Zhang et al.[25] proposed an instance-level multi-view transfer algorithm (MVTL-LM) that integrates classification loss and views consistency terms in a large-margin framework. Jing et al.[21] proposed a Multi-view Discriminant Transfer (MDT) learning approach for domain adaptation. Unlike MVTL-LM, our method operates at the feature level, i.e., it mines the correlations between views together with the domain distance measure to improve transfer.

However, in many real-world scenarios, given a target domain, there may be more than one source domain available for building classifiers. In this case, how to fully utilize multiple sources to ensure effective knowledge transfer is a crucial issue. Thus far, several approaches have been proposed for transfer learning with multiple source domains. Most of them are focused on learning weights for different domains based on the similarities between each source domain and the target domain or learning more precise classifiers from the source domain data jointly by maximizing their consensus of predictions on the target domain data.

3. OVERVIEW

3.1 Preliminaries

Definition 1 (City Grid): We divide a city into disjointed grids, assuming that placement in a grid g is uniform; each grid has several data samples and only one label that denotes whether it contains only one store (if not, we will try to use a store with more data as a label).

Definition 2 (Social View): Information aggregation index *svi* obtained by the analyses of online user review data of smart cities.

Definition 3 (Physical View): Information aggregation index *pvi* obtained through offline urban region data from various physical sensors and satellite data from smart cities.

3.2 Framework

As shown in Figure 1, our framework consists of two major components: feature extraction of the original and target cities, and transfer learning, which involves the analysis of optimal retail store placement in other cities. We retrieved massive amounts of online user review data from big cities. Through proper feature learning from social and physical views, we fed these data into our framework. Then, through multi-view discriminant transfer learning, we transferred knowledge to other cities with sparse online user review data.

Problem statement: Formally, by considering the existence of a candidate set of areas L in which a commercial enterprise is interested in placing its business, we wish to identify the optimal area l ∈ L such that a newly opened store in l will potentially attract the largest number of visitors. An area l is derived from grid g_i. We compute a score y_i for every candidate area l: the top-ranked area in terms of this score will be the optimal area for placing the new store. Our main assumption in the formulation of this task is that the Dianping score empirically observed by users can be used as a proxy for the relative popularity of a place.

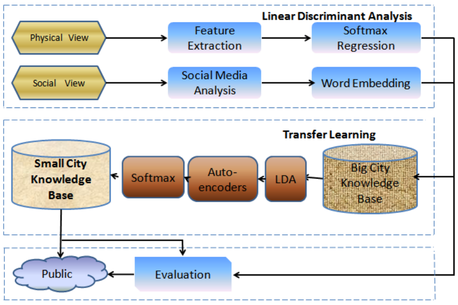


Figure 1: Multi-view discriminant transfer learning framework.

Suppose that we are given a group of source cities $\{C_{s1}, C_{s2},$ ..., $C_{sn}\}$ and a target city C_t. Each source city has a set of grids $C_{si} = \{D_s\}$, each grid has labeled source-domain data $D_s = \{(s^{(i)},p^{(j)},y^{(k)})\}$, and the target city has labeled target-domain data $D_t = \{(s^{(m)}, p^{(n)}, y^{(p)})\}$, $m << i$ and $n \approx j$, consisting of two views, where s_i and p_i are column vectors of the ith instance from the social and physical views, respectively, and y_i is its class label, $y_i \in \{0, 1, 2, ..., k\}$ (k=3 in this paper). The different class label(0,1,2,3) corresponds to the store's score. The source and target domain data follow different distributions.

Our goal is to assign appropriate class labels to the instances in the target domain. We adopt autoencoders to construct a feature mapping from an original instance to a hidden representation, and we use the source domain data jointly to train a classifier for predictions on the target domain. Eventually, we generate a final score y_i for a region g.

3.3 Map Segmentation

A road network usually consists of some major roads such as highways and ring roads that partition a city into regions[24][23]. We display the vector-based road network on a plane by performing map projection, which transforms the surface of a sphere (i.e., the Earth) into a 2D plane (we used Mercator projection in our implementation). Then, we convert the vector-based road network into a raster model by gridding the projected map. Intuitively, each pixel of the projected map image can be regarded as a grid-cell of the raster map. Consequently, the road network is converted into a binary image, e.g., after the dilation operation, the road segments are turgidly thickened. Then, we extract the skeleton of the road segments while retaining the topology structure of the original binary image. Figure 2 shows the result of the procedure described above for Beijing's entire road network. Finally, we get the grids g of cities.

4. APPROACH

[1] http://www.cs.berkeley.edu/

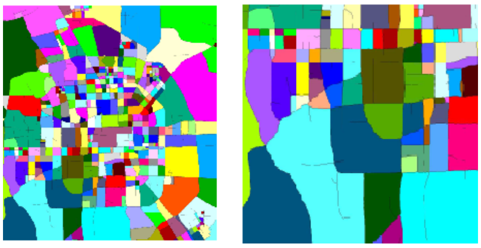


Figure 2: Segmented regions.

4.1 Model Social View

Prosperity and users' opinions of a neighborhood are two important factors determining property investment value. Recent studies have shown that a strong regional economy usually indicates high demand[12][9]. Thus, we decided to mine online user reviews collected from dianping.com.

1)Dianping Score. For each region g, we measure (1) overall satisfaction, (2) service quality, (3) environment class, and (4) consumption level r_i by mining the reviews of business venues located in $r_i, \{p : p \in P\&p \in ri\}$, where P is the set of business venues in a city.

Overall Satisfaction: For each grid g, we access the overall satisfaction of users over the neighborhood r_i. Since the overall rating of a business venue p represents the satisfaction of users, we extract the average of the overall ratings of all business venues located in r_i as a numeric score of overall satisfaction. Formally, we have
$f_i^{OS} = \frac{\sum_{p:p,q \in P\&p \in r_i} OverallRating_p}{\{p:p \in P\&p \in r_i\}}$.

Service Quality: Similarly, we compute the average service rating of business venues in r_i and express the service quality of the neighborhood as
$f_i^{SQ} = \frac{\sum_{p:p,q \in P\&p \in r_i} ServiceRating_p}{\{p:p \in P\&p \in r_i\}}$.

Environment Class: The environment class of business venues could reflect whether the neighborhood is high-class. Therefore, we extract the average environment rating as
$f_i^{EC} = \frac{\sum_{p:p,q \in P\&p \in r_i} EnvironmentRating_p}{\{p:p \in P\&p \in r_i\}}$.

Consumption Cost: The average cost of consumption behaviors in business venues can partially reflect the income and neighborhood class. We calculate the average consumption cost of business venues of a targeted neighborhood as a feature:
$f_i^{CC} = \frac{\sum_{p:p,q \in P\&p \in r_i} AverageCost_p}{\{p:p \in P\&p \in r_i\}}$.

2)Dianping Comments. We extend the existing word-embedding learning algorithm and develop five-layer neural networks for learning, as shown in Figure 3. We learn the store-specific word embedding from tweets, leveraging massive tweets as distant supervised corpora without any manual annotations. These automatically collected tweets contain noise and thus cannot be directly used as gold training data to build sentiment classifiers. However, they are sufficiently effective to provide weakly supervised signals to train store-specific word embedding.

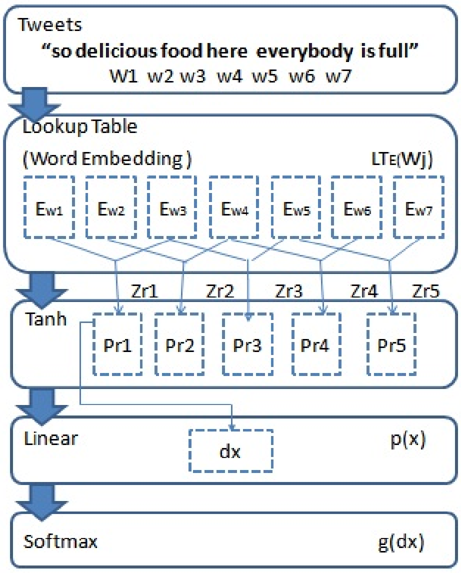


Figure 3: Structure of word embedding.

Assuming that there are K labels, we modify the dimension of the top layer in the C&W model[5] as K, and add a *softmax* layer on the top layer. The *softmax* layer is suitable for this scenario because its outputs are interpreted as conditional probabilities. Unlike the C&W model, our model does not generate any corrupted n-gram. Let $f^g(t)$, where K denotes the number of polarity labels, be the gold K-dimensional multinomial distribution of input t and $\sum_k f_k^g(t) = 1$. The cross-entropy error of the *softmax* layer is given by

$$loss_h(t) = -\sum_{k=0,1} f_k^g(t) \cdot log(f_k^h(t)), \qquad (1)$$

where $f^g(t)$ is the gold event distribution and $f^h(t)$ is the predicted event distribution.

Formally, the tweet representation is defined as
$p(x) = \frac{1}{N}\sum_{j=1}^{N} P_{\gamma j}$.

We define this probability using a softmax function:

$$g_\theta(x) = \begin{pmatrix} \frac{exp(\theta_1^T x)}{\sum_{j=1}^4 exp(\theta_j^T x)} \\ \frac{exp(\theta_2^T x)}{\sum_{j=1}^4 exp(\theta_j^T x)} \\ \frac{exp(\theta_3^T x)}{\sum_{j=1}^4 exp(\theta_j^T x)} \\ \frac{exp(\theta_4^T x)}{\sum_{j=1}^4 exp(\theta_j^T x)} \end{pmatrix},$$

where the probability of a higher-class label is the score obtained from the classifier.

4.2 Model Physical View

Recent studies have reported that different types of transit systems have different impacts on a region owing to the differences in fares, frequencies, speeds, and scopes of service. Economic information of a region can also reflect its pulse. Bus transits are slow, cheap, and mainly distributed in areas having a large number of IT and educational establishments. The price of real estate and the traffic congestion index indicate whether the facility planning is balanced. We exploit these features to uncover the implicit preferences for a neighborhood.

Bus-Related Features.: Most moderate-income residents choose buses, which are cheaper and travel at acceptable speeds, instead of taxies, which are expensive and travel at faster speeds. Because most residents in a city are from a middle-class background, bus traffic represents the majority of urban mobility. Moreover, there is a connection between a drop in grid and decreased bus mobility. Thus, we measure the arriving, departing, and transition volumes of buses in the neighborhood of each grid. Let BT denote the set of all bus trajectories of a city, each of which is denoted by a tuple < p,d >, where p is a pickup bus stop and d is a drop-off bus stop.

Bus Arriving, Departing, and Transition Volume: We extract the arriving, departing, and transition volumes of buses from smart card transactions. Formally,
$F_i^{BAV} = |< p, d > \ \ inBT : p \notin r_i \& d \in r_i|$
$F_i^{BLV} = |< p, d > \ \ inBT : p \in r_i \& d \notin r_i|$
$F_i^{BTV} = |< p, d > \ \ inBT : p \in r_i \& d \in r_i|$.

Bus Stop Density: Recent studies have reported that price premiums of up to 10% are estimated for retail stores within 300 m of a large number of bus stops. In other words, bus stop density is positively correlated to retail store value. Here, we propose an alternative approach and strategically estimate bus stop density using smart card transactions. In smart card transactions, the ticket fare of a trajectory reflects the number of bus stops in this trajectory. This is because the Public Transportation Group charges passengers according to the number of stops in each trip. Given the pick-up stop p and the drop-off stop d, the trip distance between p and d is fixed in a particular bus route. Then, the ratio of trip distance to number of bus stops implicitly suggests the average distance between two consecutive bus stops. Since the number of bus stops in a trip can be approximated by the fare, we compute the ratio of distance to fare for estimating the density of bus stops in a neighborhood. The smaller the distance-fare ratio, the higher is the bus stop density.
$F_i^{BSD} = \frac{\sum_{p \in r_i || d \ \ in r_i} dist)(p,d)/fare(p,d)}{|<p,d> \ \ inBT : p \in r_i \& d \in r_i|}$.

Smart Card Balance: The smart card balance indicates the patterns of consumption and recharge behaviors. If residents always maintain a higher balance in their smart cards, this suggests that they spend more money on bus travel. The large expense on bus travel implies that (1) residents depend on buses more than on other modes of transportation (e.g., subway, taxi), which may indicate that the affiliated neighborhood lacks subways and taxies; and (2) residents travel a long distance to work, shop, and pick up children, and thus need to maintain a high balance. In other words, this place is remote and inconvenient. Thus, we decided to extract the smart card balance as a feature. Formally, $F_i^{SCB} = \frac{\sum_{p \in r_i || d \ \ in r_i} lalance(p,d)}{|<p,d> \ \ inBT : p \in r_i \& d \in r_i|}$.

Real Estate Features: Recent studies report that real estate prices reflect the purchasing power and economic index of this region. First, we collect the historical prices of each estate, and we calculate the average estate price of the neighborhood of each grid. Formally, we have
$F_i^{RE} = \frac{\sum_{p:p,q \in P \& p \in r_i} RealEstate_p}{\{p : p \in P \& p \in r_i\}}$.

Traffic Index Features: Increased travel velocity and reduced traffic congestion should be reflected by values. We investigate the traffic index from nitrafficindex.com, which gives us as value to evaluate local traffic conditions in each grid. Formally, we have
$F_i^{RE} = \frac{\sum_{p:p,q \in P \& p \in r_i} TrafficIndex_p}{\{p : p \in P \& p \in r_i\}}$.

Competitiveness Features: We devise a feature to factor in the competitiveness of the surrounding area. Given the type of the place under prediction γ_l (e.g., Coffee Shop for Starbucks), we measure the proportion of neighboring places of the same type γ_l with respect to the total number of nearby places. Then, we rank areas in reverse order, assuming that the least competitive area is the most promising one:
$\hat{X}_l(r) = -\frac{N_{\gamma l}(l,r)}{N(l,r)}$.

However, it is worth noting that competition in the context of retail stores and marketing can have either a positive or a negative effect. One would expect that, for instance, placing a bar in an area having a large number of nightlife spots would be rewarding, as there already exists an ecosystem of related services and a large number of people are attracted to that area. However, being surrounded by competitors may also mean that existing customers will be shared.

Quality by Jensen Features: To consider spatial interactions between different place categories, we exploit the metrics defined by Jensen et al.[10]. To this end, we use the inter-category coefficients described to weight the desirability of the places observed in the area around the object, i.e., the greater the number of the places in the area that attract the object, the better is the quality of the location. More formally, we define the quality of location for a venue of type γ_l as
$\hat{X}_l(r) = \sum_{\gamma_p \in \Gamma} log(\chi_{\gamma_p -> \gamma_l}) \times (N_{\gamma_p(l,r)} - \overline{N_{\gamma_p(l,r)}})$,

where $\overline{N_{\gamma_p(l,r)}}$ denotes how many venues of type γ_p are observed on average around the places of type γ_l, Γ is the set of place types, and $\chi_{\gamma_p -> \gamma_l}$ are the inter-type attractiveness coefficients. To compute the latter, we analyze how frequently places of type γ_l are observed around γ_p on average, and we normalize that value of the expectation for a random scenario. Formally, we get
$\chi_{\gamma_p -> \gamma_l} = \frac{N - N_{\gamma_p}}{N_{\gamma_p} \times N_{\gamma_l}} \sum_p \frac{N_{\gamma_l}(p,r)}{N(p,r) - N_{\gamma_p}}$.

POIs: The category of POIs and their density in a region indicate land use as well as patterns in the region, thereby contributing to optimal placement. A POI category may even have a direct causal relation to it. Let $\sharp(i,c)$ denote the number of POIs of category $c \in C$ located in g_i, and let $\sharp(i)$ be the total number of POIs of all categories located in g_i. The entropy is defined as

$$f_i^{POI} = -\sum_{c \in C} \frac{\sharp(i,c)}{\sharp(i)} \times \log \frac{\sharp(i,c)}{\sharp(i)}. \tag{2}$$

Because these physical features are mostly related to urban infrastructure, we analyze the data through softmax regression and fit them in a single physical view[1]. The log-likelihood is given by

$$\iota(\theta) = \sum_i^m log \prod_{l=1}^4 (\frac{e^(\theta_l^T x^{(i)})}{\sum_{j=1}^4 e^(\theta_l^T x^{(i)})})^{1\{y^{(i)}=l\}}. \tag{3}$$

Now, we can obtain the maximum likelihood estimate of the parameters by maximizing $\iota(\theta)$ in terms of θ.

As shown in Eq.(4), the probability that x belongs to class k can be expressed as

$$P(y^{(i)} = k | x^{(i)}; \theta) = \frac{exp(\theta_k^T x^{(i)})}{\sum_{j=1}^4 exp(\theta_j^T x^{(i)})}, \tag{4}$$

where θ is the dimensional weight vector of physical features. Physical features are finally fitted to reveal their probability of impact on the region. Thus, we obtain two views.

4.3 Multi-view Discriminant Analysis

Diethe et al. extended Fisher's Discriminant Analysis (F-DA) to FDA2 by incorporating labeled two-view data into the Canonical Correlation Analysis (CCA) framework as follows [8][16]:

$$\max_{(w_s,w_p)} \frac{w_s^T M_w w_p}{\sqrt{w_s^T w_s} \cdot \sqrt{w_p^T M_p w_p}}, \tag{5}$$

where

$M_w = X_s^T y y^T Z_s$,

$M_s = \frac{1}{n}\sum_{i=1}^{n} (\phi(s_i) - \mu_s)(\phi(s_i) - \mu_s)^T$,

$M_p = \frac{1}{n}\sum_{i=1}^{n} (\phi(p_i) - \mu_p)(\phi(p_i) - \mu_p)^T$,

where μ_s and μ_p are the means of the source data from the two views. The numerator in Eq.(5) reflects the inter-class distance, which needs to be maximized, while the denominator reflects the intra-class distance, which should be minimized. The above optimization problem is equivalent to selecting vectors that maximize the Rayleigh quotient:

$$r = \frac{\zeta^T Q_w \zeta}{\zeta^T P \zeta}, \tag{6}$$

where $Q_w = \begin{pmatrix} 0 & M_w \\ M_w^T & 0 \end{pmatrix}, P = \begin{pmatrix} M_s & 0 \\ 0 & M_p \end{pmatrix}, and \zeta = \begin{pmatrix} w_s \\ w_p \end{pmatrix}$. Note that Q_w encodes the inter-class distance, whereas P encodes the compound information about the view-based intra-class distances. Further, ζ is an eigenvector. Such an optimization is different from FDA2 and facilitates its extension to cross-domain scenarios, which will be presented in the following subsection. For an unlabeled instance, the classification decision function is given by

$$f(s_i, p_i) = [w_s^T \phi(s_i) + w_p^T \phi(p_i) - b], \tag{7}$$

where b is the threshold.

4.4 Autoencoders

An autoencoder first maps an input instance x to a hidden representation z through an encoding mapping:

$z = h(Wx + n)$,

where h is a nonlinear activation function, $W \in R^k \times m$ is a weight matrix, and b $\in R^k \times 1$ is a bias vector. The resulting latent representation z is then mapped back to a reconstruction $\hat{x}$ through a decoding mapping:

$\hat{x} = g(W^{'} z + b^{'})$,

where g is a nonlinear activation function, $W^{'} \in R^k \times m$ is a weight matrix, and $b^{'} \in R^k \times 1$ is a bias vector. Given a set of inputs $\{x_i\}_{i=1}^{n}$, the parameters of an autoencoder are optimized by minimizing the reconstruction error as follows:

$min_{W,b,W'b'} = \sum_{i=1}^{n} ||x_i - \hat{x_i}||^2$.

4.5 Transfer Learning

The proposed optimization problem for transfer learning is formulated as follows:

$$\begin{aligned} min_{\Theta,\Theta',\{\theta_j\}} =& \epsilon(x_S, \hat{x}_S, x_T, \hat{x}_T) + \gamma\Omega(\Theta, \Theta^{'}) \\ &+ \alpha\iota(z_S, y_S; \{\theta_j\}) \end{aligned}, \tag{8}$$

where the first term in the objective is the reconstruction error of the source and target domain data, which can be written as

$$\epsilon(x_S, \hat{x}_S, x_T, \hat{x}_T) = \sum_{i=1}^{r}\sum_{i=1}^{n_j} ||x_{S_i} - \hat{x}_{S_i}||^2 + \sum_{Ti=1}^{n} ||x_i - \hat{x}_{T_i}||^2. \tag{9}$$

The second term in the objective is a regularization term on the parameters $\Theta = \{W, b\}$ and $\Theta^{'} = \{W^{'}, b^{'}\}$.

The third term represents the total loss of the softmax regression classifier over the corresponding source label data with the hidden representation. The trade-off parameters α, γ are small positive contents to balance the effect of the different terms on the overall objective.

The optimization problem is an unconstrained optimization with five types of variables to be optimized, namely $W, b, W^{'}, b^{'}, and\{\theta_j\}$. We propose the use of gradient descent methods for the solution.

Algorithm 1 Multi-view Transfer Learning with Autoencoders

```
Input:
The source dataset $D_s = \{(s^{(i)}, p^{(j)}, y^{(k)})\}$
The target dataset $D_t = \{(s^{(m)}, p^{(n)}, y^{(p)})\}$
trade-off parameters $\alpha$, $\gamma$, and the number of hidden features k.
Output:
A classifier on the target domain.
1:Initialize $W, b, W^{'}$,and $b^{'}$ by executing an autoencoder algorithm on instances of all the domains, and train $\theta_j$ on the corresponding domain data independently.
2. Fix $\theta_j$; update $W, b, W^{'}$, and $b^{'}$ alternatively.
3. Fix $W, b, W^{'}$, and $b^{'}$; update $\theta_j$.
4. If the solutions converge, construct a target classifier; otherwise, go to Step 2.
```

4.6 Target Classifier Construction

After the solutions of $W, b, W^{'}, b^{'}, andj$ are obtained, one can construct a classifier f_T in terms of $_T$ for the target domain. For any instance x_T from the target domain, which can be either from the observed unlabeled sample D_T or from an unseen data sample, the classifier f_T makes a prediction on it based on

$f_T(x_T) = \frac{1}{r}\sum_{j=1}^{r}(\theta_j^T(g(Wx_T + b)))$,

where g(x) is the classifier function of softmax regression.

.

5. EXPERIMENTS

5.1 Datasets

Table 1 lists the data sources. We extract features from smart card transactions in five cities. Each bus trip has an associated card id, time, expense, balance, route name, and pick-up and drop-off stop information (names, longitudes, and latitudes). In addition, we crawl the traffic index from nitrafficindex.com, which is open to the public. Furthermore, we crawl online business reviews from www.dianping.com, which is a site for reviewing business establishments in China. Each review includes the shop ID, name, address, latitude, longitude, consumption cost, star rating, poi category, city, environment, service, overall ratings, and comments. Finally, we crawl the estate data from www.soufun.com, which is the largest real-estate online system in China.

Table 1: Details of the datasets

Data Sources	Properties	Statics
Cities	Number of cities	5
	Number of grids	1,245
Bus Traces	Number of bus stops	9,810
	Time period	31
	Number of trips	6,543
Dianping	Number of business POIs	1472
	Number of reviews	470846
	Number of users	159302
Real Estates	Number of real estates	2,851
Traffic Index	Number of monitored regions	793
POIs	Number of POIs	3,214

5.2 Evaluation Metrics

To verify the effectiveness of our method, we compared our method with the following algorithms: (1) **MART**, a boosted tree model, specifically, a linear combination of the outputs of a set of regression trees; and (2) **RankBoost** , a boosted pairwise ranking method, which trains multiple weak rankers and combines their outputs as a final ranking. These methods use only data from the target cities.

Normalized Discounted Cumulative Gain.

The discounted cumulative gain (DCG@N) is given by

$$DCG[n] = \begin{cases} rel_1 if \quad n = \quad 1 \\ DCG[n-1] + \frac{rel_n}{log_2 n} if \quad n >= \quad 2 \end{cases}$$

Later, given the ideal discounted cumulative gain $DCG^{'}$, NDCG at the n-th position can be computed as NDCG[n]= $\frac{DCG[n]}{DCG^{'}[n]}$. The larger he value of NDCG@N, the higher is the top-N ranking accuracy.

Precision and Recall. Because we use a four-level rating system ($3 > 2 > 1 > 0$) instead of binary rating, we treat the rating 3 as a high value and ratings less than 2 as low values. Given a top-N grid list E_N sorted in descending order of the prediction values, the precision and recall are defined as Precision@N $= \frac{E_N \bigcap E_{>=2}}{N}$ and Recall@N $== \frac{E_N \bigcap E_{>=2}}{E_{>=2}}$, where $E_{>=2}$ are grids whose ratings are greater than or equal to 2.

5.3 Feature Evaluation

We provide a visualization analysis to validate the correlation between the extracted features and store values. We use a scatter-plot matrix for correlation analysis. Each non-diagonal chart in the scatter-plot matrix shows the correlation between a pair of features whose feature names are listed in the corresponding diagonal charts. Given a set of N features, there are N-choose-2 pairs of features, and thus, the same number of scatter plots. The dots represent the scores of stores and their colors represent the grades of the values. For readability, we use green > yellow > blue > red (symbol) to represent $3 > 2 > 1 > 0$ (number) in Figure 4.

In Figure 4(a), we present the correlation between social view features (overall satisfaction, service quality, environment class, consumption cost, comments) and store value. As can be seen, the green dots tend to appear at the top right corner of all the non-diagonal charts. This implies that if mobile users have higher ratings for store neighborhoods, the store values are the higher.

In Figure 4(b), we show the positive correlation between store value and bus-related features, such as the departing, arriving, and transition volumes of buses, and bus stop density, and negative correlation such as smart card balance.

Figure 4(c) shows that the traffic index has a negative correlation with store value, whereas the others have a positive correlation. Recall that by the entropy of frequency of categorized POIs, we mean the heterogenesis of POI planning. Interestingly, we observe that if the heterogenesis of functionality planning is too high or too low, these region are usually low-value region. This can be intuitively explained by the fact that people are willing to go to a place that can meet and balance the needs of their lifestyles.

The visualization results show the collectiveness of our intuitions for defining and extracting discriminative features.

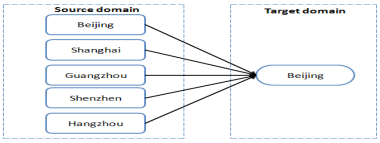
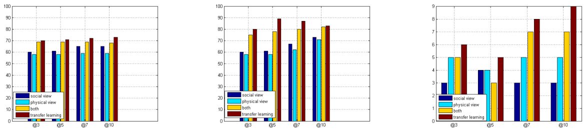


Figure 5: Source domain and target domain for transfer learning.

5.4 Model Evaluation

We use data for a single city as the baseline for our experiments. The dataset contains data from five cities in China. Each city's grid that has a retail store is annotated with a score of $\{0, 1, 2, 3\}$ based on the dianping.com score empirically observed by users. Each city is considered as a domain, and each domain contains hundreds of grids. We randomly select one of the five domains as the target domain, and all the domains serve as the source domains as shown in Figure 5 . Therefore, we can formulate four multi-source classification problems.

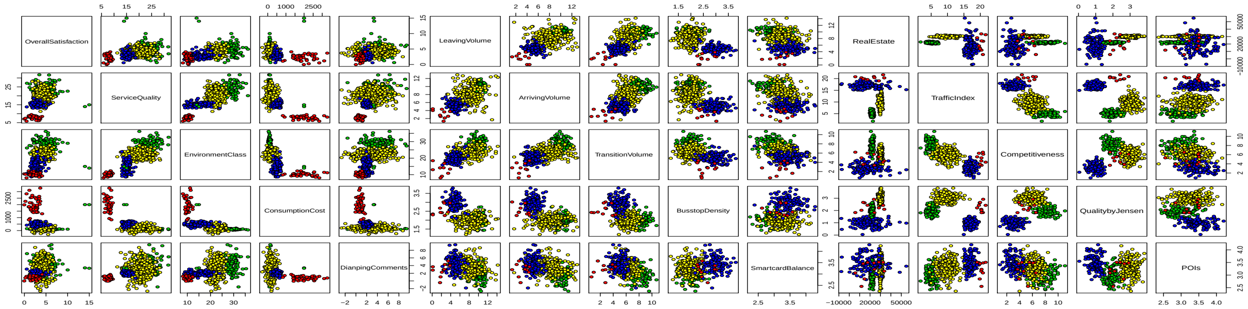
Now, considering the application scenario where the best geographic area for a new business has to be discovered, for instance, by a geo-analytics team, we would like to compare the different ranking strategies in terms of their ability to yield high-quality locations. To this end, we measure the fraction of time for which the optimal location in the predicted list R is in the top-X% of the actual popularity list R, which represents our ground truth. We refer to this metric as Accuracy@X.Note that we have used percentage instead of the absolute values (i.e., top-K) to allow for comparison across different chains.



(a) NDCG@N (b) Precision@N (c) Recall@N

Figure 6: NDCG, precision, and recall of @N for Starbucks in Beijing.

Figure 6 shows the NDCG, precision, and recall of the social view, physical view, both views, and transfer learning for Starbucks in Beijing. In all cases, we observe the performance of discriminant transfer learning outperformed the



(a) Features of social view. (b) Features of physical view (1). (c) Features of physical view (2).

Figure 4: Feature correlation analysis of social view and physical view.

other methods.

In Table 2, we present the results obtained for the NDCG@10 metric for all features across the three chains. In all cases, we observe a significant improvement with respect to the baseline. Nevertheless, by controlling the number of training grids in the cities, as shown in Figure 7, we found that the error rate decreases as the number of grids increases. In fact, the number of training samples increases with the number of grids.

Overall, transfer learning yields better results than the single-city method. Moreover, when considering more city grids in addition to the original city grids, we observed considerable improvements, highlighting the greater accuracy provided by a larger number of city grids.

Table 2: The best average NDCG@10 results of baseline and transfer learning

Cities	Starbucks	TrueKungFu	YongheKing
MART (Single City)			
Beijing	0.743	0.643	0.725
Shanghai	0.712	0.689	0.712
Hangzhou	0.576	0.611	0.691
Guangzhou	0.783	0.691	0.721
Shenzhen	0.781	0.711	0.722
RankBoost (Single City)			
Beijing	0.752	0.678	0.712
Shanghai	0.725	0.667	0.783
Hangzhou	0.723	0.575	0.724
Guangzhou	0.812	0.782	0.812
Shenzhen	0.724	0.784	0.712
Transfer Learning			
Beijing	**0.755**	**0.712**	**0.755**
Shanghai	**0.745**	**0.751**	0.783
Hangzhou	**0.755**	**0.711**	**0.752**
Guangzhou	0.810	**0.810**	**0.823**
Shenzhen	0.780	0.783	**0.723**

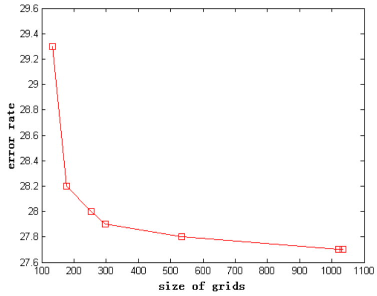


Figure 7: Sample size and error rate.

6. CONCLUSIONS

In this paper, from the perspective of a smart city, we analyzed retail store placement using four datasets observed in cities in general. Using the proposed multi-view discriminant transfer learning method, we transferred knowledge from some cities to other cities with sparse data. In addition, we tested our approach for five cities in China. The results showed that our approach is applicable to different city environments.

The proposed transfer learning algorithm may also have the same effect for some type of data sparsity, such as AQI in small cities. Thus, we hypothesize that our algorithm will succeed in transferring other urban knowledge, such as air pollution and traffic, from larger cities to smaller cities and towns with sparse data. This can be understood by analyzing the differences between cities and the rich knowledge transfer obtained from big cities.

In the future, we plan to apply our approach to other cities. Moreover, the sparsity of labeled data for machine learning remains a problem. To overcome this problem, it is necessary to use transfer learning to train sparse labeled data with abundant labeled data.

7. ACKNOWLEDGMENTS

This work is funded by NSFC61070156,YB2013120143 of Huawei and Fundamental Research Funds for the Central Universities.

8. REFERENCES

[1] L. S. Aiken, S. G. West, and S. C. Pitts. Multiple linear regression. *Handbook of psychology*, 2003.

[2] A. Bassoli, J. Brewer, K. Martin, P. Dourish, and S. Mainwaring. Underground aesthetics: Rethinking urban computing. *Pervasive Computing, IEEE*, 6(3):39–45, 2007.

[3] P. H. Calais Guerra, A. Veloso, W. Meira Jr, and V. Almeida. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2011.

[4] M. Chen, K. Q. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *Advances in neural information processing systems*, pages 2456–2464, 2011.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[6] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *Advances in neural information processing systems*, pages 353–360, 2008.

[7] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200. ACM, 2007.

[8] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor. Multiview fisher discriminant analysis. In *NIPS workshop on learning from multiple sources*, 2008.

[9] Y. Fu, Y. Ge, Y. Zheng, Z. Yao, Y. Liu, H. Xiong, and N. J. Yuan. Sparse real estate ranking with online user reviews and offline moving behaviors.

[10] P. Jensen. Network-based predictions of retail store commercial categories and optimal locations. *Physical Review E*, 74(3):035101, 2006.

[11] Y.-S. Ji, J.-J. Chen, G. Niu, L. Shang, and X.-Y. Dai. Transfer learning via multi-view principal component analysis. *Journal of Computer Science and Technology*, 26(1):81–98, 2011.

[12] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 793–801. ACM, 2013.

[13] S. Liu, Y. Liu, L. M. Ni, J. Fan, and M. Li. Towards mobility-based clustering. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 919–928. ACM, 2010.

[14] S. Liu, S. Wang, and F. Zhu. Structured learning from heterogeneous behavior for social identity linkage. 2015.

[15] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 51–62. ACM, 2014.

[16] T. Melzer, M. Reiter, and H. Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern recognition*, 36(9):1961–1971, 2003.

[17] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[18] S. D. Roy, T. Mei, W. Zeng, and S. Li. Socialtransfer: cross-domain transfer learning from social streams for media applications. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 649–658. ACM, 2012.

[19] S. Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

[20] Z. Xu and S. Sun. Multi-view transfer learning with adaboost. In *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*, pages 399–402. IEEE, 2011.

[21] P. Yang and W. Gao. Multi-view discriminant transfer learning. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1848–1854. AAAI Press, 2013.

[22] Y. Yao and G. Doretto. Boosting for transfer learning with multiple sources. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1855–1862. IEEE, 2010.

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

[24] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong. Discovering urban functional zones using latent activity trajectories. 2014.

[25] D. Zhang, J. He, Y. Liu, L. Si, and R. Lawrence. Multi-view transfer learning with a large margin approach. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1208–1216. ACM, 2011.

[26] F. Zhuang, X. Cheng, S. J. Pan, W. Yu, Q. He, and Z. Shi. Transfer learning with multiple sources via consensus regularized autoencoders. In *Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer, 2014.