Modeling of Geographic Dependencies for Real Estate Ranking

YANJIE FU, Missouri University of Science and Technology
Email: yanjiefoo@gmail.com
HUI XIONG, Rutgers University
Email: hxiong@rutgers.edu
YONG GE, University of Arizona
Email: ygestride@gmail.com
YU ZHENG, Microsoft Research
Email: yuzheng@microsoft.com
ZIJUN YAO, Rutgers University
Email: zijun.yao@rutgers.edu
ZHI-HUA ZHOU, Nanjing University
Email: zhouzh@lamda.nju.edu.cn

It is traditionally a challenge for home buyers to understand, compare and contrast the investment value of real estate. While a number of appraisal methods have been developed to value real properties, the performances of these methods have been limited by traditional data sources for real estate appraisal. With the development of new ways of collecting estate-related mobile data, there is a potential to leverage geographic dependencies of real estate for enhancing real estate appraisal. Indeed, the geographic dependencies of the investment value of an estate can be from the characteristics of its own neighborhood (individual), the values of its nearby estates (peer), and the prosperity of the affiliated latent business area (zone). To this end, in this paper, we propose a geographic method, named ClusRanking, for real estate appraisal by leveraging the mutual enforcement of ranking and clustering power. ClusRanking is able to exploit geographic individual, peer, and zone dependencies in a probabilistic ranking model. Specifically, we first extract the geographic utility of estates from geography data, estimate the neighborhood popularity of estates by mining taxicab trajectory data, and model the influence of latent business areas. Also, we fuse these three influential factors and predict real estate investment value. Moreover, we simultaneously consider individual, peer and zone dependencies, and derive an estate-specific ranking likelihood as the objective function. Furthermore, we propose an improved method named CR-ClusRanking by incorporating checkin information as a regularization term which reduces the performance volatility of real estate ranking system. Finally, we conduct a comprehensive evaluation with the real estate related data of Beijing, and the experimental results demonstrate the effectiveness of our proposed methods.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Filtering; H.2.8 [Database Management]: Database Applications—Data Mining

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Real Estate, Clustering, Ranking, Geographic Dependencies
1. INTRODUCTION

Different from market value, real estate investment value is the intrinsic long-term “worth” of an estate. It reflects the growth potential of an estate’s resale value that can be higher or lower than market value to a particular investor. In fact, a high price does not necessarily mean a high investment value, and vice versa. This characteristic motivates investors to enter real estate marketplace, seek estates with high investment value, and maximize their investment returns. Therefore, ranking estates based on investment value is important and urgent, because it can enable many applications, such as supporting decision marking for home buyers, optimizing price structure for housing brokers, and enhancing site selection for real estate developers. In the real estate industry, real estate investment value is usually quantified by the investment returns of an estate over certain holding periods, e.g., rising and falling market periods [Williams 1938]. An estate, in this paper, refers to a residential complex which has one or more buildings where each building has many condos, and we define estate return rate as the ratio of the price increase relative to the start price of a market period [Feibel 2003].

Classic estate appraisal methods, e.g., financial time series analysis and automated valuation models, typically predict the market value of an estate rather than its investment value. Traditional Learning-To-Rank (LTR) methods can rank estates by treating estates as documents and their investment values as relevance scores, but their performances are limited by lack of combining both ranking techniques of documents and geographic dependencies of estates in modeling.

With the development of new ways of collecting estate-related mobile data, there is a potential to exploit geographic dependencies of estates for enhancing estate appraisal. As a matter of fact, a large amount of estate-related mobile data, such as urban geographic data and human mobility information near estates, have been accumulated. If properly analyzed, these data can provide a source of rich intelligence for finding estates with high investment values. Specifically, we study three types of geographic dependencies, which categorize estate values from three perspectives: (1) the geographic characteristics of its own neighborhood (individual), (2) the values of its nearby estates (estate-estate peer), and (3) the values of its affiliated latent business area (estate-business zone).

First, the investment value of an estate is largely determined by the geographic characteristics of its own neighborhood. We call it individual dependency. For example, people are usually willing to pay higher prices for estates close to the best public schools. The individual dependency can be captured by correlating the estate investment values with urban geography (e.g. bus stops, subway stations, road network entries, and point of interests (POIs)) as well as human mobility patterns. Second, the estate investment value can be reflected by its nearby estates. We call it peer dependency. The peer dependency can be captured by the comparative estate analysis which is a popular method in estate appraisal and evaluates estates based on peer estate comparison. An intuitive understanding along this line is, if the surrounding estates are of high investment values, the targeted estate usually has a high value.

Third, the estate value can also be influenced by the values of its affiliated latent business area. We call it zone dependency. A business area is a self-organized region with many estates. The formation of business areas are driven by the long-term commercial activities under two mutually-enhanced effects: (1) estates tend to co-locate in
multiple centers, and thus bring human activities to those business areas; (2) prosperous business areas in return lead to more estate constructions. Hence, a prosperous business area represents a high density cluster of human activities, commercial activities, and estates. Here, we assume that each estate is affiliated with a latent business area and each business area is endowed with a value function of estate investment preferences, which measures the prosperity of the estate industry in this business area. The more prosperous the business area is, the easier we can identify a high investment-value estate from this business area.

In addition, people tend to visit popular POIs in or near the centers of business areas for leisure activities and check in to these POIs for social purposes. These checkins reflect their geographic preferences implicitly. Similar to estates, POI checkins are also distributed along multiple business areas in a city, and thus share the same multi-cluster patterns with estates. To this end, we jointly model the correlations among estates, checkins and business areas via topic modeling and treat checkin information as an additional regularization term.

In summary, the individual dependency shows that the investment value of an estate can be reflected by urban geography information and human mobility data of its neighborhood. Also, the peer dependency allows to exploit spatial autocorrelation of investment values through the comparison between the targeted estate and its peer estates. Moreover, the zone dependency allows to explore the influence of the associated latent business area of an estate. Finally, checkins help regularize zone dependency.

Based on the above, we propose a framework for estate appraisal by leveraging the mutual enforcement of ranking and clustering. This framework is able to exploit geographic individual, peer and zone dependencies, as well as checkin regularization, into a unified probabilistic ranking model. Specifically, we first extract the geographic utility from urban geography data. Then, we estimate the neighborhood popularity through spatial propagation and aggregation of passenger visit probabilities by mining taxicab trajectory data. Moreover, we model the influence of latent business areas by embedding a dynamic spatial-clustering approach into the ranking process where each business area reflects the influence on estate investment values like a spatial hidden state. After this, we fuse the three factors and estimate estate investment values. We simultaneously consider individual, peer and zone dependencies and derive a mixture likelihood objective. We also consider checkins of POIs and use them as a regularization term for the spatial clustering, enhance latent business area analysis and reduce system volatility.

In our preliminary work [Fu et al. 2014b], we only considered three geographic dependencies (i.e., individual, peer and zone) and proposed the ClusRanking method. In this paper, we further develop an improved method named CR-ClusRanking, which not only captures geographic individual, peer and zone dependencies, but also regularizes by checkins which share the similar multi-center patterns with real estates regarding spatial distribution. CR-ClusRanking analogizes <checkin pattern, estate neighborhood, business area> to <word, document, topic> and exploits topic modeling to jointly model the correlations among them, such that we are able to extract priors of business area prosperities and seamlessly integrate these priors as a regularization term into ClusRanking. Finally, we conduct a comprehensive performance evaluation on the real estate related data of Beijing and the experimental results demonstrate ClusRanking and CR-ClusRanking outperform several baselines. Moreover, CR-ClusRanking further increases ranking accuracy and reduces volatility of estate ranking system comparing to ClusRanking.

The identified geographic dependencies can be generalized from three perspectives. First, the geographic dependencies are not only applicable in the urban areas of Beijing, but also in the urban areas of many other big cities in the world as long as these
cities have the needs for the mixed-land use development. Indeed, the development of mixed land use blends a combination of land uses, such as residential, commercial, entertainment, cultural, institutional, or industrial uses, in order to physically and functionally integrate a variety of urban functions. There are many development projects with mixed-land uses in many countries, such as Shanghai, Shenzhen, and Guangzhou in China as well as Charlotte and Miami in USA. Therefore, while we use the data from Beijing for experiments, the methods developed in this paper can be helpful for the development of similar systems for other cities in the world. Second, in urban areas, in addition to real property, many geo-items (e.g., retail stores, restaurants, etc.) have similar spatial autocorrelations including individual dependency, peer dependency, and zone dependency. The identified geographic dependencies can be generalized to build geographic ranking systems for business venues on site selection. Finally, modeling different geographic dependencies can help fuse and compare geographic information from multiple sources, and thus effectively reduce the influence of untrustworthy information. Also, the identified geographic dependencies present models of ranking representations from different perspectives, and thus help reduce the risk of choosing a wrong modeling hypothesis of geographic rankings.

2. PRELIMINARIES

Real estate investment value refers to the capital growth potential of an estate in its future resale. The future resale price of an estate can be higher or lower than its current market price in terms of its investment value. Real estate investment value can be reflected by the investment return of an estate over a certain holding period (e.g., rising market or falling market). It thus highly necessitates a tool to rank estates based on their investment values for supporting investment decision marking. To this end, in this paper, we provide a focused study of ranking estates based on investment values. Formally, let \( E = \{ e_1, e_2, ..., e_I \} \) be a set of \( I \) estates, each of which is represented by all associated geographic features denoted as \( e_i \), as shown in Table I where more notations are listed. Our goal is to rank the estates in a descending order according to the investment values \( Y = \{ y_1, y_2, ..., y_I \} \).

Specifically, in this study, we refer an estate to a residential complex. A residential complex has one or more buildings, and a building has many condos. We assume each estate \( i \) has a location (i.e., latitude and longitude) and a neighborhood area (e.g., a circle with radius of 1 km), which we call an estate neighborhood depicted as the red dashed circles in Figure 2(a) and 2(b). In this way, we can correlate the statistical properties of urban geography data and human mobility data of an estate’s neighborhood to its investment value, because today location is the most important factor that determines the investment value of residential complexes in urban areas of a big city. To quantify estate investment value, we use the return rate of an estate given a rising or falling market period. The return rate is given by \( r = \frac{P_f - P_i}{P_i} \), where \( P_f \) and \( P_i \) denote the final price and the initial price of the market period, respectively.

In fact, the essential task of this problem is how to estimate the investment value (denoted as \( y_i \)) of each estate \( i \) by modeling all associated relevant information of estates in a unified way. In this paper, we consider a group of heterogenous information associated with estates, which include public transportation information (e.g., bus stop, subway, road network), POIs (e.g., restaurant and shopping mall), neighborhood popularity, and the influence of business areas.

\(^1\)http://en.wikipedia.org/wiki/Mixed-use
development
3. THE CLUSRANKING METHOD

We introduce the proposed ClusRanking method for real estate evaluation.

3.1. The Overview of ClusRanking

Assume that each estate \( i \) is endowed with an investment value \( y_i \). The proposed method for predicting \( y_i \) consists of two parts: (1) the predictive model and (2) the objective function.

To predict real estate investment value with geographic information, we formulate the predictive model as follows. The estate value is mainly affected by three influential factors:
\[
y_i = F(\gamma_i, \rho_i, \delta_i)
\]
where (1) \( \gamma_i \): the geographic utility extracted from urban geography data \( F_{geo} \); (2) \( \rho_i \): the influence of latent business area \( F_{area} \); (3) \( \delta_i \): the neighborhood popularity estimated from human mobility data \( F_{mobi} \). These three factors are important in evaluating real estate investment value, because the geographic utility, the neighborhood popularity, and the influence of business areas represent three different perspectives: (1) land uses, (2) human beings, and (3) business potential, respectively. There are other factors that could potentially influence the investment value of real estate, and they can also be incorporated in our method. Thus, our predictive model is extensible. After predicting \( y_i \), we will be able to get a ranked list of estates based on their predicted investment values, and thus each estate \( i \) is associated with an inferred rank \( \pi_i \).

To learn the parameters of the predictive model from the ground-truth ranked list of estates, we formulate a likelihood function, which simultaneously captures the geographic individual \( (Lik_{ig}) \), peer \( (Lik_{pde}) \) and zone \( (Lik_{ze}) \) dependencies. This likelihood function unifies both the prediction accuracy based on geographic data of estates and the ranking consistency of the estate ranked list. By maximizing this likelihood function, we can optimize the predictive accuracy of estate investment value and the ranking list of estates at the same time. We solve the optimization problem using a Expectation Maximization (EM) method. Figure 1 shows the framework of our method.

3.2. Modeling Real Estate Investment Value

Before introducing the overall objective function which captures the three dependencies altogether, let us first introduce how to model the investment values of estates...
with geographic information. Specifically, we will first introduce the modelings of $\gamma_i$, $\rho_i$ and $\delta_i$ separately, and then state how they are combined together.

3.2.1. Geographic Utility: $\gamma_i$. Estate values are largely determined by its geographic location. Therefore, we naturally relate the geographic utility of an estate to its location characteristics. Specifically, we identify the following transportation features shown in Figure 2(a): (1) **Bus**: the distance to the nearest bus stop, and the number of bus stops located in estate neighborhood. (2) **Subway**: the distance to the nearest subway station, and the number of subway stations located in estate neighborhood. (3) **Road Network**: the distance to the closest road network entry, and the number of road network entries (including exits).

Also, we extract and normalize the number of POIs of different categories as features shown in Figure 2(b). Specifically, for each estate, we first count the number of POIs of the category $c_i \in C$ in its neighborhood, denoted by $\#_i$. Because the scales of POI numbers vary over different POI categories (e.g., the POIs in the food category are much more than other categories), we treat the neighborhood of each estate as a document and treat each POI category as a word. We then apply the TF-IDF strategy to normalize the POI feature value of category $c_i$ as

$$P[c_i] = \frac{\#_i}{\sum_{i=1}^{\#} \log \frac{|H|}{\#_i}}.$$ 

Fig. 1. The framework of the ClusRanking method. The black plates represent the latent effects.

Fig. 2. (a) feature extraction of public transportation; (b) feature extraction of POIs and taxi GPS trajectories; (c) the function for propagating visiting probability of taxi passengers from a drop-off point to POIs near an estate. Note that a neighborhood is defined as a cell area with a radius of 1km.
Table II. Neighbourhood Profiling (a neighborhood is defined as a cell area with a radius of 1km.)

<table>
<thead>
<tr>
<th>Data</th>
<th>Feature Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transportation</td>
<td>Number of bus stop</td>
</tr>
<tr>
<td></td>
<td>Distance to bus stop</td>
</tr>
<tr>
<td></td>
<td>Number of subway station</td>
</tr>
<tr>
<td></td>
<td>Distance to subway station</td>
</tr>
<tr>
<td></td>
<td>Number of road network entries</td>
</tr>
<tr>
<td></td>
<td>Distance to road network entries</td>
</tr>
<tr>
<td>Point of interest</td>
<td>Number of POIs of different POI categories</td>
</tr>
<tr>
<td></td>
<td>(Shopping, Sports, Education, etc.)</td>
</tr>
</tbody>
</table>

where $|H|$ denotes the total number of estates, and $|\{h|c_i \in h\}$ stands for the total number of estates which have POIs of category $c_i$ appearing in their neighborhoods. The TF-IDF feature value increases proportionally to the number of POIs of category $c_i$ in the estate’s neighborhood, but is offset by the number of estates whose neighborhoods have POIs of category $c_i$.

Next, we detail how to model the geographic utility. Specifically, as shown in Table II, we first extract the above geographic features from estate neighborhoods, represent each estate as a geographic feature vector, and treat the estate feature vectors as raw representations of estates, denoted by a matrix $E$. The raw representations of estates $E$ are then learned and transformed to the meta representations $WE$ using a single-layer perceptron, where $W \in M \times N$ is a coefficient matrix. Finally, we parameterize geographic utility by a linear aggregation over transferred features in meta representation: $\gamma = qWE^T$, where $q \in 1 \times M$ are the weights of the transferred features.

The reason for using a single-layer perceptron $W$ and a relative coefficient vector $q$ is as follows. We have extracted many geographic features, which we call raw representation $E$, by feature engineering from a variety of data sources (e.g., bus data, subway data, road network data, and POI data). However, these features may come with different importance and usually are correlated and redundant. Therefore, it is difficult to decide which feature is more important and which might be useless. Thus, we are actually using the data to learn the relative importance (the coefficient vector $q$), and overcome the redundant issue of the intercorrelated geographic features.

3.2.2. Influence of Latent Business Area: $\rho$. There are two perspectives of modeling the formation of business areas. First, we can view business areas as a result of probabilistic clustering, in which each estate is softly assigned to K business areas with certain probabilities, where the sum of the probabilities over K business areas is 1. The investment values of estates are influenced by the values of K business areas. Alternatively, we can also view business areas as a result of geographic segmentation, where each estate is affiliated with a business area as long as the estate is spatially located in the business area. Then, the centers of business areas can be approximately estimated by the centers of corresponding estate clusters. Moreover, the influence of business area properties can be quantified in terms of distances between estates and the centers of business areas.

Next, we will adapt the idea of Gaussian Mixture Model to unify the above two perspectives and model the influence of latent business areas. Suppose there are K latent business areas, we first choose the business area for each estate. Naturally, we need an appropriate distribution to model how likely an estate can be drawn from a business area; that is, the possibility that an estate belongs to a business area. Since K business areas are indexed by integers (i.e., 1, 2, ..., K), we apply a multinominal distribution over K latent business areas $r \sim p(r|\eta)$, where $\eta \in 1 \times K$ can be explained as the values (prosperities of estate industry or estate investment preferences) of K business areas. Later, each estate location $l_i$ is drawn from a multivariate normal distribution:
distance between the estate location and the business area center: the influence on each estate. Assume the influence is inversely proportional to the distance between the estate location and the business area center: \( d(i, r) = \sqrt{\|\mu_r - l_i\|^2} \), the influence of \( K \) business areas over estate \( i \) is defined by an aggregate power-law weighted parametric term

\[
\rho_i = \sum_{k=1}^{K} \left( \frac{d_0}{d_0 + d(i, r_k)} \right)^{\eta_k} \frac{\eta_k}{\sum_{k=1}^{K} \eta_k},
\]

where \( d_0 \) is a parameter. We exploit the exponential function \( \left( \frac{d_0}{d_0 + d(i, r_k)} \right)^{\eta} \) to model the spatial pattern that if a residential location is closer to the center of a business area, it receives more positive influence. We use the mathematical constant \( e \) for simplicity. We can approximate the number of business areas (\( K \)) in terms of empirical knowledge about business areas in a city, which typically can be learned from Wikipedia and other public sources. Then, we further identify the best \( K \) by testing different values of \( K \) around the above approximated number.

3.2.3. Neighborhood Popularity: \( \delta \). Neighborhood popularity can affect the investment value of an estate to a certain extent. In general, in urban areas people are willing to live in a popular neighborhood. A popular neighborhood usually has lots of notable POIs, which can be measured from two perspectives: (1) POI numbers, representing the quantitative measurement; (2) POI visit probability of mobile users, representing the quality of those POIs. We propose to estimate the neighborhood popularity of a targeted estate by strategically combining POI numbers and POI visit probabilities using the taxicab GPS traces via a three-stage algorithm.

Propagating visit probability. In the first stage, given the drop-off point of a taxi trace \( d \), we model the probability of a POI \( p \) visited by the passenger as a parametric function, whose input \( x \) is the road network distance between \( d \) and \( p \):

\[
P(x) = \frac{\beta_1}{\beta_2} \cdot x \cdot \exp(1 - \frac{x}{\beta_2}).
\]

The reasons why we adopt this function are as follows. First, the above function has two mathematical properties: \( \beta_1 = \max d P(d) \) and \( \beta_2 = \arg\max d P(d) \). Figure 2(c) shows a sample plot of the probability with \( \beta_1=0.8 \) and \( \beta_2=25 \), which means the probability of visiting a POI reaches the highest value 0.8 when the distance between the POI and the drop-off point is 25 meters. Second, when \( x = 0 \), \( P(x) = 0 \). Since a taxi may not send passengers into a POI directly, the drop-off point usually is not the same as the destination. A passenger often walks a short distance to reach the destination. Third, the drop-off point usually is close to the destination. Hence, when the distance exceeds a threshold \( \beta_2 \), the probability keeps decreasing with an exponential heavy tail. With this function, we can propagate the visit probability of a passenger from the drop-off point to its surrounding POIs.

Aggregating POI-level visit probability. Given a POI \( p \), the visit probability of \( p \) is measured by summarizing all the visit probabilities propagated from all the drop-off points in taxicab trace data via \( \kappa(p) = \sum_{d \in D} P(\text{dist}(d, p)) \).

Aggregating POI-category-level visit probability. In the third stage, we first identify the POIs located in the neighborhood \( n_i \) of the \( i \)-th estate. Then, we summarize the visit probability of those POIs per category \( c_j \) and obtain the category-level aggregated visit probability as \( \phi_{ij} = \sum_{p \in c_j} \kappa(p) \). In this way, we reconstrucr the...
representation of neighborhood popularity as an aggregated visit probability vector \( \mathbf{\phi}_i = \langle \phi_{i1}, \cdots, \phi_{ij} \rangle \) over different POI categories for the \( i \)-th estate, where \( J \) is the total number of POI categories. Finally, we aggregate and normalize the popularity score as

\[
\delta_i = \frac{1}{J} \sum_{j=1}^{J} \frac{\phi_{ij}}{\max_{\phi_{ij}}}.
\]

Finally, we combine all modellings of \( \gamma_i, \rho_i \) and \( \delta_i \) together and get the overall generative process of estate investment values as shown in Table III. Specifically, we first assume there are \( K \) latent business areas in a city. Each business area is a cluster of estates. We treat \( K \) latent business areas as \( K \) spatial hidden states, each of which is endowed with a latent value \( \eta_k \) representing estate investment preference (or prosperity of estate industry) in the \( k \)-th business area. For each estate \( i \), we draw a business area \( r \) from all \( K \) business areas following a multinomial distribution: \( \text{Multi}(\eta) \). The location of estate \( l_i \) is drawn from the sampled business area \( r \). Later, given the estate location \( l_i \), we are able to identify the neighborhood area and represent estate by a geographic feature vector \( e_i \) via neighborhood profiling. We then extract geographic utility \( \gamma_i \) from \( e_i \). Moreover, we estimate the neighborhood popularity \( \delta_i \) by strategically mining the taxicab trajectory traces. Since the estate investment value depends on the value of the associated latent business area, the \( K \) business areas together show the value influence on the estate: \( \rho_i = \sum_{k=1}^{K} \left( \frac{d_0}{d_0 + d(i, r_k)} \right) e \sum_{k=1}^{K} \eta_k \), which is penalized by the distance between area centroid and estate location. After incorporating the three factors, we generate the investment value \( y_i \) of real estate \( i \). With all the estate investment values, we compile a ranked list of estates denoted as \( \Pi \).

### 3.3. Modeling Three Dependencies

We introduce how to model the geographic individual, peer and zone dependencies of estates together in a unified objective function, as shown in Figure 1. Let us denote all parameters by \( \Psi = \{ \eta, \mu, \Sigma \} \) where \( \eta \in 1 \times K \) is the relative coefficients of transferred meta features, \( W \in M \times N \) is the coefficient matrix of the single-layer perceptron, \( \eta \in 1 \times K \) is the parameters of multinormal distributions for drawing estates from \( K \) latent business areas, \( \mu \in K \times 2 \) and \( \Sigma \in K \times 2 \times 2 \) are the centers and covariances of \( K \) business areas, the hyperparameters \( \Omega = \{ \mu_q, \sigma_q^2, \mu_w, \sigma_w^2, \sigma_\rho^2 \} \) where \( \mu_q \) and \( \sigma_q^2 \) are the mean and variance of the prior Gaussian distribution for \( q \), \( \mu_w \) and \( \sigma_w^2 \) are the

---

**Table III. The generative process of ClusRanking**

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>For each estate ( i ):</td>
</tr>
<tr>
<td>1.1</td>
<td>Draw a business area ( r \sim \text{Multinomial}(\eta) ).</td>
</tr>
<tr>
<td>1.2</td>
<td>Draw a location ( l_i \sim \mathcal{N}(l_i; \mu_r, \sigma_r^2) ).</td>
</tr>
<tr>
<td>1.3</td>
<td>Generate geographic utility.</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Draw coefficient matrix of meta representation ( w_{mn} \sim \mathcal{N}(w_{mn}; \mu_w, \sigma_w^2) ).</td>
</tr>
<tr>
<td>1.3.2</td>
<td>Draw coefficient vector of geography utility ( q_m \sim \mathcal{N}(q_m; \mu_q, \sigma_q^2) ).</td>
</tr>
<tr>
<td>1.3.3</td>
<td>Estate geographic utility ( \gamma_i = q We_i ).</td>
</tr>
<tr>
<td>1.4</td>
<td>Compute influence given by latent business areas ( \rho_i = \sum_{k=1}^{K} \left( \frac{d_0}{d_0 + d(i, r_k)} \right) e \sum_{k=1}^{K} \eta_k ).</td>
</tr>
<tr>
<td>1.5</td>
<td>Compute neighborhood popularity ( \delta_i = \frac{1}{J} \sum_{j=1}^{J} \frac{\phi_{ij}}{\max_{\phi_{ij}}} ).</td>
</tr>
<tr>
<td>1.6</td>
<td>Generate the estate investment value ( y_i \sim \mathcal{N}(y_i; f_i, \sigma^2) ) where ( f_i = \gamma_i + \delta_i + \rho_i ).</td>
</tr>
<tr>
<td>2</td>
<td>Compile the ranked list ( \Pi ) of estates in terms of all ( y_i ).</td>
</tr>
</tbody>
</table>
are the mean and variance of the prior Gaussian distribution for \( W \), \( \sigma^2 \) is the variance of the Gaussian distribution for the ground-truth real estate investment values \( Y \), and the observed data collection \( D = \{ Y, \Pi, L \} \) where \( Y \), \( \Pi \) and \( L \) are the investment values, ranks and locations of \( I \) estates, respectively. For simplicity, we first assume that \( i = \pi_i = \pi_r \). In other words, the real estates in \( D \) are sorted and indexed in a descending order in terms of their investment values, which compiles descending ranks as well.

By Bayesian inference, we have the posterior probability as

\[
Pr(\Psi|D, \Omega) \propto P(D|\Psi, \Omega) P(\Psi|\Omega).
\]

We follow the commonly-used “bag of words” assumption [Blei et al. 2003], which in our setting corresponds to conditional independence of the investment value, ranking, and locations of an estate, given parameters \( \Phi \) and \( \Omega \), to approximate the objective function. The term \( P(D|\Psi, \Omega) \) is the likelihood of the observed data collection \( D \) as

\[
P(D|\Psi, \Omega) = P(\{ Y, \Pi, L \} | \Psi, \Omega) \propto P(\{ Y, L \} | \Psi, \Omega) P(\Pi|\Psi, \Omega),
\]

where \( P(\{ Y, L \} | \Psi, \Omega) \) denotes the likelihood of the observed investment values and locations of estates given the parameters. \( P(\{ Y, L \} | \Psi, \Omega) \) can be explained as to be proportional to the individual dependency \( \text{Lik}_{id} \). \( P(\Pi|\Psi, \Omega) \) denotes the likelihood of the ranking of estates given the parameter, which is proportional to the product of peer dependency \( \text{Lik}_{pd} \) and zone dependency \( \text{Lik}_{zd} \). Next, we introduce the modeling of each dependency in detail.

**Individual Dependency.** The smaller the loss, the higher \( \text{Lik}_{id} \). Specifically, we model \( \text{Lik}_{id} \) as a joint probability of the estate investment values, the estate locations, and the business areas to learn the geographic interinfluence between estate investment values and locations. As shown in Table III, we assume that each location of estate is drawn from a business area and all business areas are drawn from a Multinomial distribution. Along this line, \( \text{Lik}_{id} \) is formulated by

\[
\text{Lik}_{id} = \prod_i P(\{ y_i, l_i \} | \Psi, \Omega) = \prod_i \sum_{r_i=1}^{K} P(\{ y_i, l_i, r_i \} | \Psi, \Omega) P(r_i),
\]

where we introduce a latent variable \( R \in 1 \times I \) where each element \( r_i \) represents the latent business area assignment of estate \( i \).

**Peer and Zone Dependencies.** While directly modeling likelihood of the ranking list of estates could not comprehensively capture the spatial correlation of estate-estate and estate-business area, we model the ranking consistency by \( \text{Lik}_{pd} \) and \( \text{Lik}_{zd} \) instead. In fact, the ranked list of all the estates indeed can be encoded into a directed graph, \( G = \{ V, E \} \), with the node set \( V \) as estates and the edge set \( E \) as pairwise ranking orders. For instance, edge \( i \rightarrow h \) represents that an estate \( i \) is ranked higher than the estate \( h \). From a generative modeling aspect, edge \( i \rightarrow h \) is generated by our model through a likelihood function \( P(i \rightarrow h) \). The more valuable the estate \( i \) against estate \( h \), the larger the \( P(i \rightarrow h) \). Since an estate pair \( <i, h> \) can be located inside one business area or cross two different business areas, the edges of \( G \) then can be categorized into two sets: (1) edges intra business area which corresponds to peer dependency and (2) edges inter business area which corresponds to zone dependency.

Specifically, \( \text{Lik}_{pd} \) is defined as the ranking consistencies of estate pairs within the same business area. In other words, peer dependency captures the likelihood of the edges intra business area. Here the generative likelihood of each edge \( i \rightarrow h \) is defined as \( \text{Sigmoid}(f_i - f_h): P(i \rightarrow h) = 1 + \exp(-(f_i - f_h)), \) where \( f_i \) and \( f_h \) are the predicted
investment values of estate \(i\) and \(h\). Thus, \(L_{ikpd}\) is defined by

\[
L_{ikpd} = \prod_{i=1}^{I} \prod_{h=i+1}^{H} P(i \rightarrow h | \Psi, \Omega)^{I(r_{i}=r_{h})},
\]

(8)

where \(I(r_{i}=r_{h})\) is the indicator function with \(I(r_{i}=r_{h}) = 1\) when estate \(i\) and estate \(h\) are in the same business area (or \(r_{i}=r_{h}\)), and \(I(r_{i}=r_{h}) = 0\) otherwise.

While the peer dependency considers the estate pairs which are within the same business area, zone dependency targets the estate pairs whose elements are within two different business areas. We use the generative likelihood of edges inter business area as the zone dependency. There is investment value conformity between estate and business area: the higher prosperity of estate industry in the associated business area, the higher possibility we can draw a high-value estate. Thus, when the estate pair \(<i,h>\) is drawn from two different business areas \(<r_{i},r_{h}>\), we compare the values of the two associated business areas \((r_{i} \rightarrow r_{h})\) instead of the values of estates \((i \rightarrow h)\). Therefore, the generative likelihood of an inter-business-area edge is defined as \(\text{Sigmoid}(\eta_{r_{i}} - \eta_{r_{h}})\): \(P(i \rightarrow h) = \frac{1}{1+\exp(-(\eta_{r_{i}} - \eta_{r_{h}}))}\), where the values of \(r_{i}\) and \(r_{h}\) are represented by \(\eta_{r_{i}}\) and \(\eta_{r_{h}}\) respectively (refer to Section 3.2.3). In this way, we capture the spatial dependency between estate and business area. \(L_{ikzd}\) is then given by

\[
L_{ikzd} = \prod_{i=1}^{I} \prod_{h=i+1}^{H} P(r_{i} \rightarrow r_{h} | \Psi, \Omega)^{I(r_{i} \neq r_{h})}.
\]

(9)

Second, term \(P(\Psi | \Omega)\) is the prior of the parameters \(\Psi\):

\[
P(\Psi | \Omega) = \prod_{m=1}^{M} N(q_{m}, \mu_{q}, \sigma_{q}^{2}) \times \prod_{m=1}^{M} \prod_{n=1}^{N} N(w_{mn}, \mu_{w}, \sigma_{w}^{2}).
\]

(10)

4. THE CR-CLUSRANKING METHOD

In this section, we introduce the enhanced method, namely CR-ClusRanking, which regularizes ClusRanking with checkin information.

4.1. Check-in Information

In a city, the prosperities of business areas not only affect estate values, but also influence the decision process of site selection for leisure activities. For example, people tend to choose places in or near the centers of business areas, check in to these places, and show their geographic preferences. Similar to estates, checkins of POI also distribute along these business areas, and thus reflect the prosperities of business areas. Therefore we enhance our ClusRanking method via incorporating checkin information. Specifically, we exploit the topic modeling and model the correlations among business areas, estates and checkins as shown in Table IV. We treat each estate neighborhood as a document and each checkin pattern as a word. Topic modeling clusters real estates into multiple clusters, each of which is regarded as a business area. With the analogous topic model, we mine the prosperities of business areas. The more prosperous the
business area, the more likely we identify an high-valued estates and specific checkin
patterns. We utilize the mined prosperities of business areas as prior knowledge and
enhance the learning of ClusRanking. Before describing our CR-ClusRanking, let us
first introduce two definitions.

DEFINITION 1: (Checkin Pattern) Given a checkin event CI, a checkin pattern
CP=(CI.a, CI.b, CI.c) is a triple containing (1) checkin day (CI.a), (2) checkin hour
(CI.b), (3) POI category of the checkin place (CI.c).

DEFINITION 2: (Checkin Cuboids) A checkin cuboid CC is a 7×24×J cuboid,
where 7 is the number of day of week, 24 is the number of hours, and J is the number
of POI categories. The cell CC(a, b, c) represents the number of checkin patterns that
mobile users checkin at these pois of category c at hour b of day a.

4.2. The Description of CR-ClusRanking

We propose a two-step approach to extract the prosperities of business areas.

STEP1: Checkin pattern propagation from POIs to neighborhoods. We note
that each neighborhood is associated with a cluster of POIs, and each POI is associated
with a cluster of checkins. We first extract the checkin patterns for each POI. We then
propagate the extracted checkin patterns from POI to neighborhoods. The neighbor-
hood of each estate, i, therefore corresponds to a cluster of checkin patterns, called di.
One reason for propagating checkin patterns from POIs to neighborhoods is that the
checkin patterns associated with a single POI are usually homogeneous and incom-
plete. The spatial aggregation can better learn the prosperities of business areas in

terms of topic distribution.

STEP2: Prosperity estimation of business areas from <checkin, estate,
business area> triples. We exploit the probabilistic latent semantic analysis (PLSA),
a classic topic modeling method, for prosperity estimation. In PLSA, each document is
represented as a probability distribution over topics and each topic is represented as
a probability distribution over a number of words. The model has two latent variables
that can be inferred from the data: (1) the document-topic distributions, and (2) the

To apply PLSA, we first treat each checkin pattern as a word. A check-in event con-
tains the information of user ID, POI name, latitude and longitude, POI category, and
time stamp, and thus a checkin event is usually unique. However, unlike a checkin
event, a checkin pattern, according to DEFINITION 1, is defined as a triple: (checkin
day, checkin hour, checkin POI category), which is a summarization of a checkin event.

Although a checkin event is unique, a checkin pattern could be shared by different
checkin events. For example, if a mobile user U1 check in to a shushi bar (POI cate-
gory: restaurant) in Queens, New York City at 10pm, Thursday, the checkin pattern
of U1 is denoted as (Thursday, 10pm, Restaurant). In addition, as stated in STEP1,
we associate all these checkin patterns to a nearby estate neighborhood, once their
checkin points are located within the circle area of the estate with a radius of 1 km.

Since a checkin pattern is regarded as a word, and an estate neighborhood is regarded
a bag of checkin patterns, we treat each estate neighborhood as a document. Back to
the above example, suppose another mobile user U2, rather than user U1, also check
in to a buffet (POI category: restaurant) at 10pm, Thursday. The checkin pattern of
U2 is denoted as (Thursday, 10pm, Restaurant). As can be seen, although U2's checkin
event is different from U1's checkin event, U2's checkin pattern is the same with U1's
checkin pattern, which will be represented by the same word in PLSA. Because the
locations of U1's checkin event is different from the location of U2's checkin event, the

ACM Transactions on Knowledge Discovery from Data, Vol. V, No. N, Article xx, Publication date: x 2016.
two checkin events are associated to two different estate neighborhoods. Therefore, a "word" (i.e., checkin pattern) can be shared across different "documents" (i.e., estate neighborhoods). Finally, because a business area contains many estates, we treat each business area as a topic, which is a cluster of documents in topic modeling. We can use an analogous PLSA model to jointly capture the correlations among business areas (i.e., topics), estate neighborhoods (i.e., documents), and user checkin patterns (i.e., words): (i) different business areas generate different checkin patterns and corresponding frequencies; (ii) different checkin patterns reflect different functions and properties of business areas. In this way, we can learn latent business areas and the corresponding properties (densities) of business areas from real estate data and user checkin data.

We build the analogous PLSA model with the generative process as follows. Let \( w, d \) and \( z \) denote checkin pattern, estate neighborhood and business area respectively only. Topic distribution \( P(z) \) describes the prosperity of business area; topic distribution of document, \( P(w|d) \), illustrates the influence of prosperity of business area \( z \) for estate \( d \); \( P(w|z) \) represents the likelihood we identify checkin pattern \( w \) from business area \( z \).

- For the document \( d_i \), given an estate \( i \)
  - For each word \( w_{d,x} \) in document \( d_i \):
    a. Draw a topic \( z_{d,x} \sim Mult(\theta_{d,i}) \)
    b. Draw a word \( w_{d,x} \sim Mult(\phi_{z_{d,x}}) \)

where \( x \) represents the index of word \( w \) in document \( d \).

The likelihood of the aggregated PLSA is

\[
Pr = \prod_d \prod_w P(d, w)^{n(d, w)}
\]

where \( n(d, w) \) is the number of \( w \) in document \( d \). \( P(d, w) \) is traditionally defined as \( P(d) \sum_z P(w | z)P(z | d) \). By Bayesian rule, we can reorder \( P(d, w) \) as \( P(d) \sum_z P(w | z)P(z | d) = \sum_z P(w | z)P(d | z)P(z) \). The objective of reordering is to extract \( P(z) \) which illustrates the prosperities of \( K \) business areas learned from the intercorrelation among estates, checkins and business areas. The log-likelihood of the observed data is \( O = \sum_d \sum_w n(d, w) \log (\sum_z P(w | z)P(d | z)P(z)) \).

The parameter estimation of the model can be implemented by an Expectation-Maximization (EM) algorithm given the checkin and real estate data. Specifically, we can infer

\[
P(z) = \frac{\sum_d \sum_w n(d, w)P(z | w, d)}{\sum_d \sum_w n(d, w)},
\]

where \( P(z | w, d) = \frac{P(z | d)P(w | z)}{\sum_z P(z | d)P(w | z)} \). Here \( P(z) \) is of our interest. In ClusRanking, we place an non-informative conjugate Dirichlet prior to update the multinomial distribution (\( \eta \)) of \( K \) business areas. However, with the extracted \( P(z) \), we can further incorporate \( P(z) \) and place an informative conjugate Dirichlet distribution on \( \eta \), so that \( \eta \) is regularized by \( P(z) \) which will be detailed in the following.

5. MODEL INFERENCE

With the formulated posterior probability, the learning objective is to find the optimal estimation of the parameters \( \Psi \) that maximize the posterior. There are latent variables in our proposed method (note that \( r_i \in R \) is the latent assignment of business area for estate \( i \)). Therefore, we use EM mixed with a sampling algorithm. EM mixed with a sampling algorithm is also called the Monte Carlo EM method (MCEM). The work in [Fort and Moulines 2003] established the convergence of the MCEM algorithm. The algorithm iteratively updates the parameters by mutual enhancement between Geo-clustering and estate ranking. The Geo-clustering updates the latent busi-
ness areas based on locations and the three geographic dependencies; estate ranking learns the estate scores and generate a ranked list.

**E-Step.** In the E-step, we iteratively draw latent business area assignments for all real estates. For each estate $i$, we treat its latent business area $r$ as a latent variable, which is drawn from the posterior of $r$ in terms of the complete likelihood: 

$$r \sim P\left(l_i \mid r, \Psi^{(t)}\right) P\left(Y, \Pi \mid r, \Psi^{(t)}\right) P\left(r \mid \eta^{(t)}\right),$$

where

$$P\left(Y, \Pi \mid r, \Psi^{(t)}\right) = P(y_i \mid f_i, \sigma^2) \prod_{h=i+1}^l P(h \mid r, \Psi^{(t)})^{I(r_i = r_h)},$$

$$\prod_{h=i+1}^l P(h \mid r, \Psi^{(t)})^{I(r_i \neq r_h)}.$$ 

Here the latent business area assignment of real estate $c_i$ is updated by three effects: (1) $P(r \mid \eta^{(t)})$ updates business area assignment in terms of the prosperity distribution of multiple business areas; (2) $P\left(l_i \mid r, \Psi^{(t)}\right)$ is the location emission probability given the latent business area as a hidden spatial state. (3) $P\left(Y, \Pi \mid r, \Psi^{(t)}\right)$ updates business area assignment by both prediction accuracy and ranking consistency because $Y$ and $\Pi$ respectively represent locational ratings and ranking orders (refer to 6). When the latent business area assignment of each estate is updated, we further update the neighborhood popularity $\delta_i = \frac{1}{l} \sum_{j=1}^l \frac{\phi_{ij}}{\max_{i \in r(\phi_{ij})}}$, because the normalization term is conditional on the updated business area $r_i$.

**M-Step.** In the M-step, we maximize the log likelihood of the model given the business area assignments $R$ fixed in the E-step. Since business area assignments are known, we can update $\mu_r, \Sigma_r, \eta$ directly from the samples as follows:

$$\mu_r = \frac{1}{\#(i,r)} \sum_{i=1}^l I(r_i = r) l_i,$$

$$\Sigma_r = \frac{1}{\#(i,r) - 1} \sum_{i=1}^l I(r_i = r) \left( (l_i - \mu_r)^T (l_i - \mu_r) \right),$$

where $\#(i,r)$ is the number of real states assigned to region $r$.

**The update rule of ClusRanking:** For ClusRanking, by imposing a conjugate non-informative Dirichlet prior $\text{Dir}(\gamma)$, we update $\eta^{(t+1)}$ by

$$\eta_r^{(t+1)} = \frac{C_r^{(t+1)} + \gamma r}{C^{(t+1)} + \sum_r \gamma r},$$

where $C_r = \sum_{i \in r} y_i$, $C = \sum y_i$, and $\gamma = \frac{1}{K}$. Here we initially assume the influences of K business areas are all the same, that is, $\gamma$ is a non-informative prior.

**The update rule of CR-ClusRanking:** For CR-ClusRanking, we impose a conjugate informative Dirichlet prior with the learned $P(z)$, we update $\eta_r^{(t+1)}$ by

$$\eta_r^{(t+1)} = \frac{C_r^{(t+1)} + P(z_r)}{C^{(t+1)} + \sum_r P(z_r)},$$

where $r$ is the index of business area. Different from ClusRanking, we use the learned prosperities of business areas to initialize $\eta$. Note that the centers ($\mu$) and estate in-
vestment values ($\eta$) of latent business areas are updated, so updated is the influence of latent business areas $\rho_i = \sum_{k=1}^{K} \left( \sum_{c=1}^{c} \sum_{k=1}^{K} \eta_k \right)$.

After updating the parameters $\{\eta, \mu, \Sigma\}$ and latent business area assignments $R = \{r_i\}_{i=1}^{I}$, we update $\Psi^{(t+1)}$ that maximizes the log of posterior

$$L(q, W|R^{(t+1)}, D) = \sum_{i=1}^{I} \left[ -\frac{1}{2} \ln \sigma_q^2 - \frac{(q_m - \mu_q)^2}{2\sigma_q^2} \right] + \sum_{m=1}^{M} \sum_{n=1}^{N} \left[ -\frac{1}{2} \ln \sigma_w^2 - \frac{(w_{mn} - \mu_w)^2}{2\sigma_w^2} \right]$$

(17)

We apply a gradient descent method to update $q, W$ through $q_m^{t+1} = q_m^t - \epsilon \frac{\partial (-L)}{\partial q_m}$ and $w_{mn}^{t+1} = w_{mn}^t - \epsilon \frac{\partial (-L)}{\partial w_{mn}}$. Note that the latent business area assignment of each estate $r_i$ are updated in each E-step, so updated are the neighborhood popularity and the influence of latent business areas. Then the model parameters are updated upon the new business area assignments.

6. RANKING PREDICTION

After parameters $\Psi$ are estimated via maximizing the posterior probability, which essentially captures both prediction accuracy of estate investment value and the ranking consistence of estates, we will obtain the learned model for investment value of estate, i.e., $\mathbb{E}(y_i|q, e_i) = \gamma_i + \delta_i + \rho_i$ given a rising or falling market period. For a new coming estate $k$, we may predict its investment value accordingly. The larger the $\mathbb{E}(y_k|q, e_k)$ is, the higher investment value it has. With the predicted investment values for all new estates, we are able to compile a ranking list of those estate.

7. EXPERIMENTAL RESULTS

In this section, we provide an empirical evaluation of the performances of the proposed ClusRanking and CR-ClusRanking methods on real-world estate data.

7.1. Data Description

Table V shows five data sources we used in the experiments. The transportation dataset includes the data about the bus system, the subway system, and the road network in Beijing, China. Also, we extracted POI features from the Beijing POI dataset. Moreover, mobility patterns are extracted from the taxi GPS traces. In Beijing, taxi traffic contributes more than 12 percent of the total traffic, and thus reflects a significant portion of human mobility [Yuan et al. 2012]. Furthermore, we collected the Beijing checkin data from www.weibo.com, a Chinese version of Twitter. Finally, we crawled the Beijing estate data 2 from www.soufun.com, which is the largest real-estate online system in China. Although the data sources donot cover the entire rising and falling market periods, we used the collected data to approximate the geo-mobile information of these time windows missing real world data. This is because (1) urban infrastructures of a city change slowly in a small time period, and (2) spatiotemporal patterns of human mobility have periodicity.

We use real estate return rates to measure the investment value of an estate over rising and falling markets. The reasons are as follows. Real estate investment value is a

2http://goo.gl/0R2F15
Table V. Statistics of the experimental data.

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Properties</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of real estates</td>
<td>2,851</td>
</tr>
<tr>
<td></td>
<td>Size of bounding box (km)</td>
<td>40°x40</td>
</tr>
<tr>
<td></td>
<td>Time period of transactions</td>
<td>04/2011 - 09/2012</td>
</tr>
<tr>
<td>Real estates</td>
<td>Return rate over rising market</td>
<td>-0.76 (min)/0.02 (mean)/1.25 (max)</td>
</tr>
<tr>
<td></td>
<td>Return rate over falling market</td>
<td>-0.76 (min)/-0.04 (mean)/2.2 (max)</td>
</tr>
<tr>
<td></td>
<td>Distance to bus stop (meter)</td>
<td>0.08 (min)/199.12 (median)/4973.85 (max)</td>
</tr>
<tr>
<td></td>
<td>Number of bus stops within 1 km</td>
<td>0 (min)/16 (median)/823 (max)</td>
</tr>
<tr>
<td></td>
<td>Distance to subway station (meter)</td>
<td>24.47 (min)/1124.95 (median)/10103.92 (max)</td>
</tr>
<tr>
<td></td>
<td>Number of subway stations within 1 km</td>
<td>0 (min)/2 (median)/13 (max)</td>
</tr>
<tr>
<td></td>
<td>Distance to road network entry (meter)</td>
<td>1.92 (min)/16.06 (median)/283.27 (max)</td>
</tr>
<tr>
<td></td>
<td>Number of road network entries within 1 km</td>
<td>3 (min)/527 (median)/1082 (max)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bus stop(2011)</td>
<td>Number of bus stop</td>
</tr>
<tr>
<td></td>
<td>Subway(2011)</td>
<td>Number of subway station</td>
</tr>
<tr>
<td></td>
<td>Road networks (2011)</td>
<td>Number of road segments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total length(km)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Percentage of major roads</td>
</tr>
<tr>
<td></td>
<td>POIs</td>
<td>Number of POIs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of categories</td>
</tr>
<tr>
<td></td>
<td>Taxi Trajectories</td>
<td>Number of taxis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Effective days</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of trips</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of GPS points</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total distance(km)</td>
</tr>
<tr>
<td></td>
<td>Checkins</td>
<td>Number of poi category</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of checkin events</td>
</tr>
</tbody>
</table>

property’s intrinsic long-term worth. Property investment value is usually represented by the return an investment asset would have to yield. In other words, the underlying drivers for property investment ratings are the “dividends” or “capital gains” over a certain holding period (rising and falling market periods in this study). It is true there are many macro and micro factors that could potentially impact real estate investment returns. But no matter how many factors there, the impact of these factors can be finally reflected by the real world market performances of real properties in rising and falling market periods.

To prepare the benchmark investment values of estates (denoted by Y) for training data, we first calculated the return rate of each estate during a given market period. We then sorted the return rates of all the estates in a descending order. Finally, we clustered them into five clusters using variance based top-down hierarchical clustering. In this way, we segmented the estates into five ordered value categories (i.e., 4 > 3 > 2 > 1 > 0, the higher the better). By discretizing estate return rates into five categories, we can understand estate investment potentials and reduce the noise led by the small fluctuations in return rates.

Finally, a list of estates, each of which with the extracted features and investment values, were split into two data sets in terms of the falling market period (from Jul. 2011 to Feb. 2012) and the rising market period (from Feb. 2012 to Sep. 2012) as shown in Figure 3.

7.2. Evaluation Metrics

We aim to build ranking systems for real estate on evaluating investment values, as people usually care more about retrieving top-N estates of highest investment values for investment decision making. Therefore, to show the effectiveness of the proposed ranking systems, we used the following metrics (i.e., NDCG, Recall, Tau) for measuring
Fig. 3. The rising market period and the falling market period in Beijing.

ranking accuracy. Also, to evaluate the performance volatility of the ranking systems, we use variance as an evaluation metric.

**Normalized Discounted Cumulative Gain. (NDCG)** We considered NDCG@N as an standard evaluation criterion. The discounted cumulative gain (DCG) metric measures ranking quality over top $N$ estates on the result list by assuming that high-value estates should appear earlier in the ranked list. Specifically, the DCG@N is given by

$$DCG[n] = \begin{cases} rel_1 & \text{if } n = 1 \\ DCG[n-1] + \frac{rel_n}{\log_2{n}}, & \text{if } n > 2 \end{cases}$$  \hspace{1cm} (18)$$

where $rel_n$ is the ground-truth estate investment value of the $n$-th estate in the predicted estate ranking list. Later, given the ideal DCG@N $DCG'$ (i.e., the maximum value of DCG@Ns), NDCG@N can be computed as $\text{NDCG}[n] = \frac{DCG[n]}{DCG'[n]}$. The larger NDCG@N is, the higher top-N ranking accuracy is.

**Recall.** Since we used a five-level rating system ($4 > 3 > 2 > 1 > 0$) instead of binary rating, we treated the rating $\geq 3$ as “high-value” and the rating $< 3$ as “low-value”. Given a top-N estate list $E_N$ sorted in a descending order of the prediction values, recall is defined as $\text{Recall}@N = \frac{|E_N \cap E_{\geq 3}|}{|E_{\geq 3}|}$, where $E_{\geq 3}$ are the estates whose ratings are greater or equal to $3$.

**Kendall’s Tau Coefficient.** Kendall’s Tau Coefficient (or Tau for short) measures the overall ranking accuracy. Let us assume that each estate $i$ is associated with a benchmark score $y_i$ and a predicted score $f_i$. Then, for an estate pair $<i,j>$, $<i,j>$ is said to be concordant, if both $y_i > y_j$ and $f_i > f_j$ or if both $y_i < y_j$ and $f_i < f_j$. Also, $<i,j>$ is said to be discordant, if both $y_i < y_j$ and $f_i > f_j$ or if both $y_i > y_j$ and $f_i < f_j$. Tau is given by $\text{Tau} = \frac{\#\text{concordant} - \#\text{discordant}}{\#\text{concordant} + \#\text{discordant}}$.

**Performance Volatility.** We measured the performance volatility by variance based on a sample of ranking quality measurements (e.g., recall, tau), $X$. The variance is given by $\text{Var} = \frac{\sum_{x \in X} (x - \bar{x})^2}{n-1}$ where $n = |X|$.

7.3. Baseline Algorithms

Since our work is related to Learning-To-Rank, we compared our method against the following algorithms. To show the effectiveness of the proposed method, we compared the ranking accuracies of our methods:

1. **ClusRanking:** ClusRanking combines the ideas of pairwise learning to rank and Gaussian Mixture Model, and jointly models the geographic individual, peer, and zone dependencies for real estate ranking, based on investment values.

2. **CR-ClusRanking:** CR-ClusRanking is an improved version of the ClusRanking model. CR-ClusRanking not only captures geographic individual, peer and zone dependencies, but also integrates check-in information for regularization. CR-ClusRanking analogizes $<$check-in pattern, estate neighborhood, business area$>$ to $<$word, document, topic$>$ and extracts business area prosperities as priors by topic modeling.
against following baseline algorithms:

(3) **MART** [Friedman 2001]: MART is a boosted tree model in which the output of the model is a linear combination of the outputs of a set of regression trees. MART is a class of boosting algorithms that may be viewed as performing gradient descent in function space, using regression trees.

(4) **RankBoost** [Freund et al. 2003]: The basic idea of RankBoost is to formalize learning to rank as a problem of binary classification on instance pairs, and then to adopt boosting approach. Like all boosting algorithms, RankBoost trains one weak ranker at each round of iteration, and combines these weak rankers as the final ranking function. After each round, the document pairs are re-weighted: it decreases the weight of correctly ranked pairs and increases the weight of wrongly ranked pairs.

(5) **Coordinate Ascent** [Metzler and Croft 2007]: Coordinate Ascent uses a loss function called the domination loss. Coordinate Ascent extends the loss by incorporating margin requirements over pairs of instances and enables the usage of multi-valued feedback. Coordinate Ascent devises a simple yet effective coordinate descent algorithm that is guaranteed to converge to the unique optimal solution.

(6) **ListNet** [Cao et al. 2007]: ListNet introduces two probability models, respectively referred to as permutation probability and top-k probability, to define a listwise loss function for learning. Neural Network and Gradient Descent are then employed as model and algorithm in the learning method.

Besides, we also compared our methods with traditional spatial autoregressive regression based method as follows.

(7) **SAR** [Wall 2004]: SAR (Spatial Autoregressive Regression) is a well-known time-series approach which predicts housing price based on its price history by combining geographic dimensions (i.e., geographic points or geographic areas). We first predicted the price of each real estate using the SAR model and then ranked these real estates according to their predicted investment returns.

Specifically, we extracted all the defined features from urban geographic data and taxi trajectory data by using the spatial R-tree and grid indexes, and used a KNN-based method to impute the values of missing features. Also, we used Mallet[^3] and applied topic modeling to extract checkin mobility patterns. Then, we fed all the features along with real estate investment ratings into ClusRanking and CR-ClusRanking, as well as baseline methods for training. We randomly divided the data into 80% for training and 20% for testing, trained the model, and calculated the Tau, NDCG and Recall values of the model predictions over test data. We repeated this process for five times, and extracted the variance values for Section 7.5 and the average values for other sections. For the baseline algorithms, we used RankLib[^4]. The parameters of baseline methods were set up based on the recommended settings, which usually can help achieve good and stable results according to the authors of baseline algorithms as follows. We set the number of trees = 1000, the number of leaves = 10, the number of threshold candidates = 256, and the learning rate = 0.1 for MART. For RankBoost, we set the number of iteration = 300, the number of threshold candidates = 10. Regarding Coordinate Ascent, we set step base = 0.05, step scale = 2.0, tolerance = 0.001, and slack = 0.001. For Spatial Autoregressive Regression, we used the spdep package in R. The parameters of ClusRanking and CR-ClusRanking were set up based on grid search mixed with empirical investigation into data distribution. For ClusRanking model, we set $\beta_1=0.8$ and $\beta_2=25$. We set $d_0 = 1$ and $d(i, r_k)$ is computed based on degree (°) instead of mile or km for simplicity. We set latent business areas $K=10$ and initialized

[^3]: http://mallet.cs.umass.edu/
[^4]: http://sourceforge.net/p/lemur/wiki/RankLib/
The mean and covariance of the locations of each business area by K-means clustering. We also set \( \eta = \frac{1}{K}, \mu_q = \mu_w = 0, \sigma_q = \sigma_w = \sigma = 35 \) and \( M=3 \) for hyperparameters. For CR-ClusRanking model, we also set topic number as \( K = 10 \), and \( \eta = P(z) \) rather than \( \eta = \frac{1}{K} \) for hyperparameter. Finally, we set the stopping criteria of our methods as maximum number of iterations greater than 200 or relative tolerance of likelihood \( |\text{lik}_t - \text{lik}_{t-1}| \) less than \( 1 \times 10^{-4} \).

The proposed methods performed major tasks offline without the requirement of processing and learning data in real time, therefore efficiency is not a big concern. Table VI provides the average running time for major steps using the data described in Table V. The experiments were performed on a x64 machine with Intel 2.60GHz dual-core CPU and 24GB RAM running Microsoft Windows 7.

<table>
<thead>
<tr>
<th>Procedures</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Estate Value Grading</td>
<td>720ms</td>
</tr>
<tr>
<td>Checkin mobility pattern extraction</td>
<td>48min</td>
</tr>
<tr>
<td>Topic modeling</td>
<td>13min</td>
</tr>
<tr>
<td>Transportation feature extraction (per estate)</td>
<td>1.07s</td>
</tr>
<tr>
<td>POI-related feature extraction (per estate)</td>
<td>1.66s</td>
</tr>
<tr>
<td>Neighborhood popularity extraction (overall)</td>
<td>286min</td>
</tr>
<tr>
<td>Ranking model training (200 epoch)</td>
<td>228min</td>
</tr>
</tbody>
</table>

### 7.4. Overall Performances

We provide a performance comparison on the rising and falling market datasets in terms of Tau, NDCG, and Recall in order to validate the effectiveness of ClusRanking.
and CR-ClusRanking. We used the similar parameter setting of Section 7.3 and all the extracted geographic features for training and testing.

**Results and Analysis.** In rising market, Figure 4 shows the ClusRanking and CR-ClusRanking methods outperform the baseline algorithms in the evaluation of Tau, NDCG, and Recall. For example, the ClusRanking and CR-ClusRanking methods achieve 0.3428617 and 0.3788716 in Tau; the NDCGs of the ClusRanking and CR-ClusRanking methods range from 0.75 to 0.87. Comparing to the ClusRanking method, the CR-ClusRanking method fuses checkin regularization, and offers a slight increase. For example, the CR-ClusRanking method shows 10.5% increase in terms of Tau, 2.4% increase in terms of NDCG@10 and 6.7% increase in terms of Recall@3 comparing to the ClusRanking method. In falling market, Figure 5(a) presents the CR-ClusRanking method achieves higher Tau values than those of the ClusRanking method and the baseline algorithms. Figure 5(b) and 5(c) show that although the ranking accuracies of the ClusRanking and CR-ClusRanking methods are very close to each other, they perform better than other LTR algorithms and Spatial Autoregressive Regression. In both the rising and falling markets, although SAR achieves a relative good performance in Tau, which indicates overall ranking accuracy, comparing to other baseline LTR algorithms, the ClusRanking families outperform SAR in top-N ranking and overall ranking. This might be because SAR incorporates geographic information while does not combine the modeling techniques of rankings. The results yield several findings: (1) we can effectively improve top-k ranking by considering individual, peer and zone geographic dependencies which structurally and geographically describe real estate ranking objective; (2) checkin regularization improves both overall ranking and top-k ranking, as it regularizes the comparison of estates inter (i.e., peer dependency) and intra (i.e., zone dependency) business areas.

7.5. Study of Performance Volatility

We report the performance volatilities of the CR-ClusRanking and ClusRanking methods and validate the superiority of regularization with checkins. Aside from using the similar parameter setting of Section 7.3 for CR-ClusRanking and ClusRanking, we fed all the extracted geographic features and used the variance as defined in Section 7.2 to measure the performance volatility of our ranking systems.

**Results and Analysis.** Table VII presents the variances of Tau, NDCG and Recall values on rising market data. As can be seen, the CR-ClusRanking method achieves much lower variances of Tau, NDCG and Tau comparing to the ClusRanking method. In falling market, the variances of Tau, NDCG and Recall of CR-ClusRanking are consistently lower than those of ClusRanking as shown in Table VIII.

We can draw two findings from the above results: (1) we can learn business area prosperities via topic modeling by treating business areas and checkin patterns as documents and words; (2) we can impose checkin regularization on the ClusRanking method as prior and reduce the performance volatility led by the variance of data processing and model learning. These findings are important as they enable us to jointly and seamless model individual, peer, zone dependencies as well as checkin regularization by exploiting multi-source information in a unified model.

7.6. Study of Geographic Dependencies

To further demonstrate the effectiveness of three geographic dependencies and checkin regularization, we designed three internal competing methods in terms of variants of posterior likelihood: (1) **Individual Dependency (ID)**, in which we only consider the individual dependency as the objective function. (2) **Peer Dependency (PD)**, in which we only consider the peer dependency as the objective function. (3) **Peer Dependency + Zone Dependency (PD+ZD)**, in which we consider the combination of peer and zone de-
Table VII. The performance volatility on the rising market.

<table>
<thead>
<tr>
<th>Metric</th>
<th>ClusRanking</th>
<th>CR-ClusRanking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tau</td>
<td>0.0013743</td>
<td>0.00000248</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.00451</td>
<td>0.002232</td>
</tr>
<tr>
<td>Recall@1</td>
<td>0.00000499</td>
<td>0.00000455</td>
</tr>
<tr>
<td>Recall@3</td>
<td>0.0000150</td>
<td>0.00000458</td>
</tr>
<tr>
<td>Recall@10</td>
<td>0.0000421658</td>
<td>0.000000036</td>
</tr>
</tbody>
</table>

Table VIII. The performance volatility on the falling market.

<table>
<thead>
<tr>
<th>Metric</th>
<th>ClusRanking</th>
<th>CR-ClusRanking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tau</td>
<td>0.00003003</td>
<td>0.00000677</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.00055170</td>
<td>0.00000014</td>
</tr>
<tr>
<td>Recall@1</td>
<td>0.0000000121</td>
<td>0.0000000027</td>
</tr>
<tr>
<td>Recall@3</td>
<td>0.000008221</td>
<td>0.000000011</td>
</tr>
<tr>
<td>Recall@10</td>
<td>0.0000080312</td>
<td>0.0000000683</td>
</tr>
</tbody>
</table>

pendencies as the objective function. Later, we compared these three methods with our two proposed methods: ClusRanking, in which we consider individual, peer, and zone dependencies simultaneously; CR-ClusRanking, in which we consider both three dependencies and checkin regularization. In the experiments, we used all the extracted geographic features, as well as the similar parameter settings of Section 7.3.

Results and Analysis. First, Figure 6(a) and Figure 6(b) show the comparison of Tau and NDCG in rising market respectively. Overall, the ClusRanking and CR-ClusRanking methods achieve good performances, yet there are dependencies which are substantially better than others in some cases. For example, considering both peer and zone dependencies can enhance the top-k ranking accuracy but degrade the overall ranking accuracy comparing to the individual dependency only. This might be because the peer and zone dependencies better capture the ranking consistency of estates than the individual dependency, as the individual dependency indeed models the loss between predicted and ground-truth investment values, rather than ranking consistency.

Second, Figure 7(a) and Figure 7(b) show the comparison of Tau and NDCG in falling market respectively. It is clear that the ClusRanking and CR-ClusRanking methods outperform ID, PD and PD+ZD. We also observe that the top-k ranking performances of CR-ClusRanking and ClusRanking are very close to each other. However, in the comparison of overall ranking, CR-ClusRanking is higher than ClusRanking.

In summary, we observe that (1) the top-k ranking performances of CR-ClusRanking and ClusRanking are similar; (2) the overall ranking performances of CR-ClusRanking are higher than ClusRanking. These results justify the spatial autocorrelation (i.e., geographic individual, peer and zone dependencies) of estate investment values. Also, we validate that (1) the three geography dependencies can enhance top-k ranking; (2) checkin regularization can improve overall ranking and reduces system volatility.

7.7. Study of Influential Factors and Geographic Features

We first study the importance of the three influential factors: (i) geographic utility, (ii) neighborhood popularity, (iii) influence of business area prosperities. Specifically, we extracted the values of the three factors for each estate from the learned ClusRanking model. Then, we fed the three factors along with the ground-true real estate investment ratings into a random forest model for training. With the trained random forest model, we extracted the importances of the three factors, which are measured by the total decrease in node impurities from splitting on the variable, averaged over all trees. Table 7.7 presents the importance ranking: geographic utility > influence of
business areas > neighborhood popularity in both rising and falling markets. The results illustrate that residents in Beijing care more about land uses of a neighborhood than prosperities of nearby business areas.

In addition, we study the effectiveness of different geographic feature sets (i.e., subway, bus stop, POI, road network and combination) with the ClusRanking method over the rising and falling markets. In the experiments, we used the similar parameter settings of Section 7.3 for ClusRanking.

**Rising Market Data.** Figure 8(a) shows the performance comparison of the five feature sets in terms of Tau: combination > road network > bus stop, subway and POI. Figure 8(b) presents the NDCG@N (N=3, 5, 7 respectively) of different feature sets. As can be seen, the NDCG@3, NDCG@5, NDCG@7 of the combination of all the four feature sets are 0.81, 0.78, and 0.82 respectively, which obviously outperform the other four individual feature sets. In Figure 8(c), the combination of all geographic features consistently yields the highest recalls. The POI features achieve the second best performance, and outperform the features of bus stop, subway, and road networks. This is might be because POIs encode city infrastructure, urban function planning, as well as visiting preference, while the features of bus stop, subway, and road networks only represent one aspect.

**Falling Market Data.** Figure 9(a) shows a comparison of the five feature sets on Tau: combination > road network > bus stop, subway and POI, which is consistent with the results on the rising market data. Figure 9(b) presents that the combination of
Fig. 8. Performance comparison of different geographic features on rising market data.

Fig. 9. Performance comparison of different geographic features on falling market data.

Fig. 10. The NDCG@N of CR-ClusRanking over different K values. Note that K represents the number of business areas in a city.

the four feature sets outperforms individual feature sets in terms of NDCG. In Figure 9(c), the feature combination strategy consistently performs better than individual feature sets with respect to Recall.

In general, the feature fusion is better, because the data from a single source could contain noisy and unreliable information. By fusing and comparing information from multiple sources, we effectively reduce the influence of untrustworthy information. The results validate the effectiveness of using multiple information fusion.

7.8. Study of Latent Business Areas

We present the parameter sensitivity of K (i.e., the number of business areas) in terms of the CR-ClusRanking method. Figure 10 reports the NDCG@N of CR-ClusRanking over different K values on both the rising and falling market data. Generally speaking, the K value can affect the ranking accuracies to certain extent. In
Fig. 11. A comparison of boundaries and investment return ratings of the learned business areas within the Beijing Fifth Ring (K=10). With K-means, three clusters are outside the Beijing Fifth Ring and thus disappear.

particular, Figure 10(a) shows that in overall the setting with K as 10 achieves the highest NDCG@Ns (N=3, 5, 7, 10) in rising market. Also, from Figure 10(b), we observe that the setting with K=10 consistently and generally outperforms other settings such as K=3, K=5, K=15. Therefore, we set K as 10 in our experiments. The results can be explained as: according to a local business blog (e.g., http://goo.gl/MTWfgs.), Beijing has ten traditional representative business areas. These business areas are: Zhongguancun area, the Asian Games and the Olympic Games area, Wangjing area, Xidan area, Wangfujing area, Central Business District(CBD) area, Wukesong area, Xizhimen area, Dongzhimen area. The Veronoi visualization of the results with K=10 can help us compare the geographic segmentation to the traditional business areas of Beijing, and thus can provide a good interpretation for geographic market segmentation.

With this setting, our model also provides a unique understanding of the latent business areas of Beijing from the perspective of real estate. Figure 11 shows our method, learned from the data of urban geography, human mobility and real estate, is more reasonable than K-means, which simply cluster the estates by location information (i.e., latitude and longitude). For instance, in Figure 11(b), NO.4 area, named Zhongguancun, is the Chinese Silicon Valley and is famous for high-tech companies. This area is a high density cluster of human mobility, residential complexes and POIs. However, in Figure 11(a), the Zhongguancun area is improperly separated into NO.3 and NO.4 area by K-means. Another example is the NO.2 and NO.8 areas, named Wangjing and CBD respectively, in Figure 11(b). Wangjing is a quick-growing residential sub-center with easy-access transportation and luxury apartments. CBD is the Center Business District with numerous financial business offices, culture media companies and high-end enterprise information services. However, in Figure 11(a), Wangjing and CBD are improperly united into NO.2 area by K-means. The side bars of Figure 11(a) and Figure 11(b) show the indexes of business areas and their corresponding average ratings of estate investment returns. These ratings show our methods can better perform geographic segmentation for real estate market and estimate prosperities of different business areas comparing to K-means.

8. RELATED WORK

Traditional research on estate appraisal are based on financial real estate theory, typically constructing an explicit index of real estate prices [Bailey et al. 1963; Krainer and Wei 2004]. Some studies rely on financial time series analysis by inspecting the trend, periodicity and volatility of real estate prices [Chaitra H. Nagaraja and Zhao 2009; Zhou and Haurin 2010]. More studies are conducted from an econometric aspect, for example, hedonic methods [Taylor 2003] and repeat sales methods [Shiller 1991]
Recent works in [Downie and Robson 2007] study the automated valuation models, which aggregate and analyze physical characteristics and sales prices of comparable properties to provide property valuations. More recent studies [Lam 1996; Kontrimas and Verikas 2011; Pace 1998] shift to computational estate appraisal and apply general additive models, SVMs, multilayer perceptron and ensemble methods to evaluate estate market value. Most of the above methods estimate sale prices rather than investment values.

Also, our work is related to LTR. There are three categories of LTR methods: point-wise, pair-wise and list-wise [Trotman 2005]. Particularly, the pair-wise methods [Freund et al. 2003; Burges et al. 2007] consider each pair of objects as an example, and try to classify the example as positive when the first object in the pair is more preferable to the second object. The work in [Shi et al. 2012] further unifies both rating error and ranking error as objective function to enhance Top-K recommendation. There are also studies that improve ranking performance by semi-supervised learning through exploiting the disagreement between two learners [Zhou et al. 2006] or combining supervised and unsupervised ranking models [Li et al. 2009]. Besides, our work aggregates geographic dependencies as modeling hypothesis, and thus related to the work in [Ruthven and Lalmas 2002] which uses evidence theory to aggregate aspects of information use in information retrieval, e.g., aggregating different representations of rankings. The combination schemes in ensemble method reduce the risk of choosing a wrong hypothesis or a wrong local optimal by combining the hypotheses [Zhou 2012].

Moreover, our work has a connection with recent studies of exploring the geographic influence for POI recommendation. The works in [Cheng et al. 2012; Fu et al. 2014a] consider the multi-center of user checkin patterns and apply a static pre-clustering method to extract the influence of geographic proximity in choosing a POI. The work in [Liu et al. 2013] exploits multi-center user mobility and embeds a POI clustering method into probabilistic matrix factorization.

Furthermore, in the field of heterogeneous information network (HIN) analysis, the works in [Sun et al. 2009a; Sun et al. 2009b] exploit both clustering and ranking for clustering multiple types of objects in HIN. Our methods are different the two works, because our methods exploit the clustering power for enhancing the global ranking of estates based on investment value, whereas the works in [Sun et al. 2009a; Sun et al. 2009b] exploit local ranking for improving the clustering in HIN.

Finally, our work is related to studies of urban computing with mobile, geography and mobility data [Zheng et al. 2014]. For example, the work in [Gonzalez et al. 2008] finds that human mobility patterns show high degree of spatial and temporal regularity. That is, individuals more likely repeat the spatiotemporal mobility pattern and visit the frequent locations in certain frequent time slots. The work in [Song et al. 2010] reports human mobility has a high predictability of 95%. Yuan et al. discovers regional functions of a city using POIs and taxi traces by topic modeling [Yuan et al. 2012]. The work in [Zheng 2015] introduces the methodologies of cross-domain data fusion, including fusing heterogeneous urban geography and human mobility data. The work in [Karamshuk et al. 2013] selects the optimal sites for retail stores by mining Foursquare data.

9. CONCLUSION

In this paper, we proposed ClusRanking and CR-ClusRanking methods for ranking estates based on their investment values. Specifically, our methods have the ability in capturing the geographic individual, peer, and zone dependencies by exploiting various estate related data. Also, our method has two advantages. First, for predictive modeling, we established a hierarchical generative structure to capture both explicit factors (i.e., geographic utility and neighborhood popularity) and latent influences (e.g., the
influence of latent business area) based on the estate data. This generative structure
profiles, filters, aggregates and fuses multi-source information to predict estate invest-
mestone values. It helps to take advantage of rich estate-related data sources. Second, in
the learning framework, we leveraged the mutual enforcement of ranking and clus-
tering. In addition, we simultaneously considered three dependencies and construct
an estate-specific ranking likelihood as the objective function for enhancing model
learning. Furthermore, we introduced checkin information for regularization. The ex-
perimental study demonstrates the effectiveness of our methods on real-world estate-
related data over several alternative methods.

Acknowledgment: Z.-H. Zhou was partially supported by the National Science Foun-
dation of China (61333014) and the Collaborative Innovation Center of Novel Software
Technology and Industrialization, Nanjing 210023, China. H. Xiong was partially sup-
ported by the 111 Project (B14020). H. Xiong was partially supported by the 111 Project
(B14020), NSFC (71329201) and the Rutgers 2015 Chancellors Seed Grant Program.
Finally, the contact author is Professor Hui Xiong (hxiong@rutgers.edu).

REFERENCES


David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. the Journal of machine
Learning research 3 (2003), 993–1022.

com/apps/pubs/default.aspx?id=68133

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise ap-
proach to listwise approach. In ICML’07.


Chen Cheng, Haiqin Yang, Irwin King, and Michael R Lyu. 2012. Fused Matrix Factorization with Geo-
graphical and Social Influence in Location-Based Social Networks. In AAAI’12.


Yanjie Fu, Bin Liu, Yong Ge, Zijun Yao, and Hui Xiong. 2014a. User Preference Learning With Multiple
Information Fusion for Restaurant Recommendation. In SDM’14.

Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. 2014b. Exploiting Geographic


Geo-spotting: Mining Online Location-based Services for Optimal Retail Store Placement. In KDD ‘13.

Vilius Kontrimas and Antanas Verikas. 2011. The mass appraisal of the real estate by computational intel-


Research-The Cutting Edge (1996).

and Management (2009).

ACM Transactions on Knowledge Discovery from Data, Vol. V, No. N, Article xx, Publication date: x 2016.


Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu. 2009a. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. ACM, 565–576.

Yizhou Sun, Yintao Yu, and Jiawei Han. 2009b. Ranking-based clustering of heterogeneous information networks with star network schema. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 797–806.


