Estimating Potential Customers Anywhere and Anytime Based on Location-Based Social Networks

Hsun-Ping Hsieh[1,2,3(✉)], Cheng-Te Li[1,2,3], and Shou-De Lin[1,2,3]

[1] Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan
{d98944006,sdlin}@csie.ntu.edu.tw, ctli@citi.sinica.edu.tw
[2] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[3] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

Abstract. Acquiring the knowledge about the volume of customers for places and time of interest has several benefits such as determining the locations of new retail stores and planning advertising strategies. This paper aims to estimate the *number of potential customers* of arbitrary query locations and any time of interest in modern urban areas. Our idea is to consider existing established stores as a kind of *sensors* because the near-by human activities of the retail stores characterize the geographical properties, mobility patterns, and social behaviors of the target customers. To tackle the task based on store sensors, we develop a method called *Potential Customer Estimator* (PCE), which models the spatial and temporal correlation between existing stores and query locations using geographical, mobility, and features on location-based social networks. Experiments conducted on NYC Foursquare and Gowalla data, with three popular retail stores, Starbucks, McDonald's, and Dunkin' Donuts exhibit superior results over state-of-the-art approaches.

Keywords: Customer prediction · Retail stores · Location-based social network · Check-in data · Store sensor

1 Introduction

Modern big cities, such as New York City, London, Paris, and Taipei, are densely and crowded areas, where not only million of people live but also a great number of business established. As time proceeds, people move around in such urban areas in either a periodic or unpredicted manner. Various kinds of retail stores (e.g. Starbucks, McDonald's and Dunkin' Donuts) usually choose the locations possessing higher potential to attract more customers to construct new venues expecting more people can bring more revenue. In other words, the number of potential customers becomes one of the most important factor for business to

A. Appice et al. (Eds.): ECML PKDD 2015, Part II, LNAI 9285, pp. 576–592, 2015.
DOI: 10.1007/978-3-319-23525-7_35

determine their geographical placement or launch campaign events. It would be very practical and useful to acquire the knowledge about where and when a particular business can attract more consumers or audience.

In this paper, we aim to estimate the potential customers of an arbitrary location and at a given time in an urban area. By referring the number of potential customers as the number of people visited there, we propose to exploit the check-in records, place information, and social connections in location-based social networks (e.g. Foursquare and Gowalla) for estimating potential customers. The central idea is to consider existing stores of the target retail business (e.g. Starbucks) as a kind of *sensors* to estimate its potential customers at other locations without stores at any time. We use Figure 1 to elaborate our idea of estimating potential customers anywhere and anytime. Take the stores of Starbucks in New York City as examples, as marked by red circles. From location-based services, we might know the historical customers (i.e., the number of check-ins) of each store month by month, illustrated by the histogram of check-in numbers in terms of months. Now Starbucks would like to construct a new store or hold a marketing campaign, with two arbitrary locations in mind, as labeled as A and B. The problem is to estimate the number of potential customers of such two locations month by month so that it is possible to acquire the knowledge which one can bring more profit when a new business or event is launched. Given the potential customers over time, Starbucks can further understand which months are more profitable.

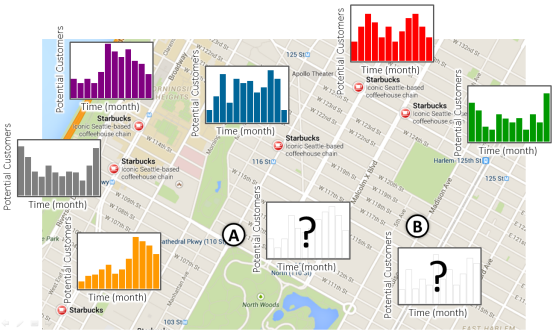

Fig. 1. An example for estimating potential customers anywhere and anytime.

Estimating the number of potential customers for arbitrary locations over time is a challenging problem. The characteristics of a location's geo-spatial neighborhood is usually one of the major factors that determine the potential customers. Such geo-spatial characteristics include population, spatial density, traffic flows, competitiveness (i.e., number of the same category of retail chain), how people interact and transit between different categories of venues, and the structure of underlying social connections in the near-by area. One major challenge of this task lies in how to model the complex composition of venues and

various moving behaviors of people in its geo-neighborhood. On the other hand, the number of potential customers of a location might change and evolve over time. Both the temporal factors of periodic growth and decline (e.g. high vs. slack seasons and weekdays vs. weekend) and special activities (e.g. anniversary and seasonal discount campaign) can affect the customer numbers. Consequently, there might not exist regular patterns to be used to predict the potential customers of a location over time. This work tries to bring such temporal impact into our estimation model.

Given some locations in a city, certain time periods of interest, and a set of stores of a target retail chain (e.g. Starbucks) that already has venues established with historical check-in data, our goal is to estimate the number of potential customers for the given locations at the designated time periods (e.g. weeks or months). To deal with this problem, we devise a model called *Potential Customer Estimator* (PCE), whose idea is three-fold. First, we construct a *Correlation Graph* (CG), which is a *multi-layer* graph, to represent the spatial and temporal correlation between existing stores and the query locations. We investigate three categories of features, *geographical*, *mobility*, and *social*, to model the correlation between locations in *CG*. Second, since different features have different effects on the estimation target, we estimate the *location correlation* separately by investigating the predictability of each feature. The correlation values derived are represented as edge weights in *CG*. Third, based on the *CG* with location correlation values on edges, we develop a *Customer Inference Algorithm* (CIA), which iteratively adjusts the estimated number of potential customers of the query locations till convergence.

2 Related Work

Investigating Location Popularity. The most relevant study is Geo-Spotting [8], which is to identify the popular locations for optimal retail store placement. Nevertheless, there are two differences. First, they formulate the task as a *ranking* problem: ranking areas such that popular areas are at the top of the list. However, we aim to estimate the exact number of potential customers, which might be more useful to calculate the potential profit of the placement. Second, while what Geo-Spotting considers is the *overall* popularity (i.e., the accumulated number of check-ins), ours is capable of estimating the number of location check-ins for a particular week, month, and season, which can be regarded as a kind of reflection of the weekly, monthly, and seasonal revenues, to facilitate the advertising strategy for the retail chain. Though the studies of Li et al. [11] investigate the common characteristics of popular locations, Kisilevich et al. [9] analyze the geo-spatial properties of attractive areas, Tiwari and Kaushik [20] design a new popularity measure based on user category, visiting frequency, and stay time, they do not make the prediction of future popularity. For other relevant work, Fu et al. [5] propose to rank the residential real estate based on investment values by mining the opinions from online user reviews and offline human mobility. Chen et al. [2] use the road network data to find locations to set up new servers such that the cost of clients being served by nearest

servers is minimized. Liu et al. [13] leverage the technique of matrix factorization to recommend locations by modeling the geographical characteristics of their neighborhoods. In addition, Hsieh et al. [6] develop a graph-based model to infer miss sensor values through learning the correlation between heterogeneous features and air quality values.

Human Mobility Prediction. Human mobility prediction is to predict the next locations that the user might visit before. Monreale et al. [16] predict the next location of a moving object with an assumption: people tend to follow common paths. With mined frequent trajectory patterns capturing common paths, they construct a decision tree-like structure, T-pattern Tree, as a predictor of the next location. Ying et al. [21] leverage the semantic information, which describes the activities (in the form of tags and types) of locations. Given the recent moves of a user, they compute the matching score geographically and semantically between mined frequent sub-trajectories and the given moves to find the the next location. Sadilek et al. [18] predict the most likely location of a user at any time, given the historical trajectories of his/her friends. They use the discrete dynamic Bayesian network to model the motion patterns of users from their friends.

3 Problem Statement

We define the number of potential customers, followed by the problem definition.

Definition 1: Number of Potential Customers. The *number of potential customers* $pc(v)$ of a location $v \in L$ is the number of check-ins performed at v, where $pc(v)$ is a positive integer, and L is the set of locations in the check-in data. In addition, the *number of potential customers* of location v at time t_i, denoted by $pc(v(t_i))$, is the number of potential customers derived from a certain time period t_i (e.g. a week or a month).

We use Foursquare check-in data in the work. We denote the maximum number of potential customers in a check-in data to be pc_{max}. Note that throughout this paper we use terms "number of potential customers", "number of check-ins", and "popularity" interchangeably.

Problem Definition: Estimating Potential Customer Number Anywhere & Anytime. Given a target retail chain and a set of its stores geographically established in the city, with the historical check-in data of the stores in time periods $T = \langle t_1, t_2, ..., t_n \rangle$, the set L of all venues in the city, the underlying social network $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ among people ($\mathbb{V}$ is the set of users and $\mathbb{E}$ is the set of social relationships between $\mathbb{V}$), an arbitrary query locations v in the city ($v \notin L$), the goal is to estimate the number of potential customers $pc(v(t_i))$ of location v in each time period $t_i \in T$.

4 Dataset

We aim to estimate the number of potential customers of query location by utilizing the check-in and venue data from the most well-known location-based service Foursquare[1] and the commonly-used location-based social network data Gowalla[2]. Since Foursquare had been launched in 2009, the volume of users, check-in records, and venue information are accumulated rapidly. Up to the end of 2013, there are 45 million users, 5 billion check-ins, and 60 million venues. Although Foursquare does not allow developers to directly access the check-in data, they allow users to share their check-ins publicly on Twitter[3]. Therefore, with the help of GEO-SPOTTING [8], we have the check-in data from Twitter and the venue data from Foursquare. To have adequate data for the experiments, we focus on New York City where Foursquare was launched and thus has significantly more users than any city in the world. The collected data in New York City contains 47,581 geo-tagged venues and 4,337,663 check-ins in a period of ten months (December 2010 to September 2011), i.e., forty weeks in total. Note that this data subset of NYC accounts for approximately 55% of all venues collected. As for the Gowalla location-based social network, which is collected by Cho et al. [3], there are totally 196,591 users, 950,327 social connections between users, and 6,442,890 check-in records collected from February 2009 to October 2010.

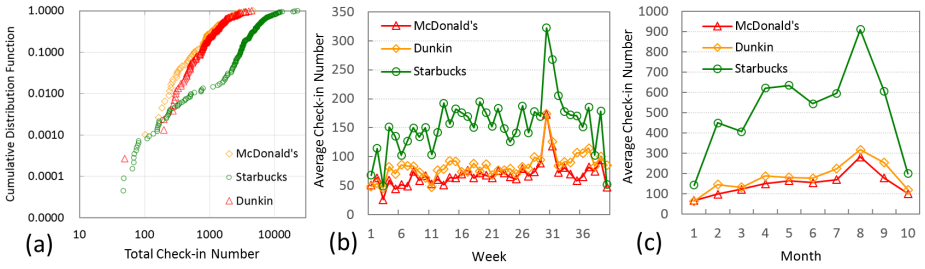

Fig. 2. Data Statistics: (a) Cumulative distribution function (CDF) of total check-ins per store for three retail chains. (b) The average check-ins of stores over weeks. (c) The average check-in of stores over months.

We target at the stores of three popular retail chains, Starbucks (SB), McDonald's (MC), and Dunkin' Donuts (DD). The statistics of each retail chain is reported as – the number of stores: 245, 89, and 149 for SB, MC, and DD; the total number of check-ins: 1,051,398, 100,520, and 187,704 for SB, MC, and DD respectively. The cumulative distribution of check-ins are also shown

[1] https://foursquare.com/
[2] https://snap.stanford.edu/data/loc-gowalla.html
[3] https://twitter.com/

in Figure 2(a). We can find that the check-in patterns of coffee shops (e.g. Starbucks) are different from those of fast food restaurants (e.g. McDonald's). Starbucks has the most number of stores as well as the most number of check-ins, in which its average number of check-in per store is almost four times than the other two chains. In addition, about 60% of Starbucks stores have check-in numbers higher than 3,000, which is significantly more than the other two as well. We believe it is because the time people take to stay in coffee shops is usually longer than that in fast food restaurants, and longer staying time would lead to higher possibility of performing check-ins. Since our goal is to estimate the number of potential customers, i.e., the evolution of check-in numbers over time, in Figure 2(b) and 2(c) we report the average number of check-ins per store over time in terms of weeks and months respectively. In general the check-in behaviors of three chains are different, except for a burst in the thirty week and in the eighth month. The average potential customer numbers of weeks fluctuate more significantly, comparing to those of months. These statistics show the difficulty of estimating potential customers.

5 Potential Customer Estimator (PCE)

5.1 Geographical, Mobility, and Social Features

We consider the following features to estimate potential customers. We calculate the features from the set of venues $N(v) = \{u : dist(u, v) < r\}$ in the near-by area with a disk of radius r centered at location v to be estimated, where $dist(u, v)$ is the geographical distance between venue u and location v. Note that we refer a venue to be a place that some kind of business has been established. We choose the radius r to be 200 meters in default according to the optimal neighborhood size suggested by the urban planning community [14]. There are three main categories of features. The first is the geographical features (GF), which describe the category distribution and the geographical interactions between venues. The specific feature items include:

- *Density* is the number of venues in the geographical neighborhood $N(v)$ of location v. The formation definition of density of location v is given by: $Density(v) = |\{u \in L : dist(u, v) < r\}|$, where L is the set of all venues in the data.
- *Neighbor Entropy* measures the heterogeneity of venue categories in $N(v)$. By denoting the set of venues with category c_i in the neighborhood of location v as $N_{c_i}(v)$ and the entire set of venue categories as C, the neighbor entropy can be defined as: $NbrEntropy(v) = -\sum_{c_i \in C} \frac{|N_{c_i}(v)|}{|N(v)|} \cdot \log \frac{|N_{c_i}(v)|}{|N(v)|}$.
- *Competitiveness* is the proportion of venues whose categories are the same as the category of the target store (e.g. "fast food restaurant" for McDonald's). Given the category of location v, denoted by c_v, its competitiveness is given by: $Compete(v) = -\frac{|N_{c_v}(v)|}{|N(v)|}$. Locations with lower competitiveness scores tend to be promising ones.

– *Attractiveness* is to capture the deployment and interactions between venue categories. If a location of a certain venue category can attract more locations with other venue categories in its neigborhood, such location is said to be more attractive. The attractiveness of location v is defined as: $Attract(v) = \sum_{c_i \in C} \log(\kappa_{c_i \to c_v}) \cdot (|N_{c_i}(v)| - |N'_{c_i}(v)|)$, where $|N'_{c_i}(v)|$ is the average number of neighboring locations of category c_i in the neighborhood of all the locations of category c_v. In addition, $\kappa_{c_i \to c_v}$ denotes the *inter-category coefficient* from category c_i to c_v. Such inter-category coefficient can be defined as: $\kappa_{c_i \to c_v} = \frac{|N|-|N_{c_i}|}{|N_{c_i}| \cdot |N_{c_v}|} \sum_{u \in L} \frac{|N_{c_v}(u)|}{|N(u)|-|N_{c_u}(u)|}$, where N is the entire set of locations in the dataset, N_{c_u} is the set of near-by locations of c_u.

The second category is the mobility features, which aim to model how users move and transit between venues. The specific feature items include:

– *Area Popularity* is the total number of check-ins for venues in $N(v)$. The formal definition of area popularity for location v is given by: $AreaPop(v) = |\{(CI(u) \in M : dist(u, v) < r\}|$, where $CI(u)$ denotes the set of check-in records at location u and M is the set of all check-ins in the data.
– *Transition Density* is the density of transitions between venues within $N(v)$. By denoting the set of consecutive check-in transitions between each pair of locations x and y as $TS((x, y) \in TS)$, the measure of transition density is defined as: $TransDensity(v) = |\{(x, y) \in TS : dist(x, v) < r \wedge dist(y, v) < r\}|$.
– *Incoming Flow* estimates the transitions from venues outside $N(v)$ to those in $N(v)$. The incoming flow score of location v is given by: $InFlow(v) = |\{(x, y) \in TS : dist(x, v) > r \wedge dist(y, v) < r\}|$.
– *Transition Attractiveness* is designed to estimate the probability of transitions between all other types of venues and venues of the same type as the target store. That says, assume people prefer to travel from locations of category c_u to locations of category c_v, if the near-by locations $u \in N(v)$ of location v can gather higher check-in numbers, then the transition attractiveness of location v tends to be high. The transition attractiveness is given by: $TransAttract(v) = \sum_{u \in N(v)} \rho_{c_u \to c_v} \cdot M_u$, where M_u is the set of check-ins at location u, and $\rho_{c_u \to c_v}$ is the probability of transitions from category c_u to category c_v. Such inter-category transition probability can be defined by the average percentage of all the check-ins from c_u to c_v: $\rho_{c_u \to c_v} = |\{(x, y) \in TS : x = u \wedge c_y = c_v\}| \cdot \frac{1}{|M_u|}$.

For the mobility features, we further consider two feature sets according to the time periods used to compute the feature values: based on the current time period to be estimated and based on the cumulative time periods from past to now. Therefore, we have two mobility feature sets: temporal mobility features (TMF), and cumulative mobility features (CMF). The third category is the social features, which characterize the social interactions for users who had ever visited the near-by area of location v. The specific feature items include:

– *Cohesiveness* is to model the structure connectivity of the graph $\mathbb{G}[N(v)]$ induced by users who had ever visit $N(v)$. This is designed to characterize the extent of cohesion or separation for people who live or visit locations within

$N(v)$. We employ the density and the clustering coefficient of $\mathbb{G}[N(v)]$ as the feature values. In addition, we also compute the number of components in $\mathbb{G}[N(v)]$ to be another indicator of cohesiveness.

- *Social Groups* estimates the number of groups of potential customers on the location-based social network $\mathbb{G}$. We consider a *community* as a group of potential customers, and calculate the number of communities for people who had ever visited $N(v)$ as the feature value. The *Louvian method* [1] is used for community detection.
- *Network Centrality* measures the importance of users on the location-based social network $\mathbb{G}$. We compute the values of degree, closeness, betweenness, PageRank, and SimRank, and consider the maximum, minimum, and average scores over users who had ever visit $N(v)$ as feature values.
- *Geo-Social Metrics* aim to quantify the geo-social influence of a user within the near-by area of location v. We employ four well-known geo-social metrics: *spatial degree centrality* [12], *spatial closeness centrality* [12], *node novelty* [19], and *geographic clustering coefficient* [19]. We compute the maximum, minimum, and average values over users who had ever visited $N(v)$.

5.2 Correlation Graph

We construct the *Correlation graph* to model the spatial and temporal correlations between existing and query locations. What follows first defines and elaborates the correlation graph, and then describes how to exploit the features extracted above to derive the *correlation* between locations as edge weights.

Definition: Correlation Graph (CG). A CG is a multi-layer weighted connected graph $G = \langle G^{t_1}, G^{t_2}, ..., G^{t_n} \rangle$, in which t is the total number of layers for time periods $t_1, t_2, ..., t_n$, and $G^{t_i} = (V, E, W)$ is the layer graph in at the t_i-th time period, where V is the set of locations, E is the set of edges between locations, and $W = W^{t_i} + W^{t_i,t_j}$ is the matrix representing edge weights, where W^{t_i} and W^{t_i,t_j} are edge weights learned from nodes within time period t_i and across time periods t_i and $t_j (i \neq j)$ respectively. The node set V consists of (a) existing locations of retail stores whose potential customer numbers have known, denoted by labeled nodes $V_\bullet$, and (b) the query locations, denoted by unlabeled nodes $V_\circ$, where $V = V_\bullet \cup V_\circ$. Each labeled node $v_\bullet \in V_\bullet$ is associated with its potential customer numbers $pc(v_\bullet(t_i))$. The edge set E also consists of two parts: the set of edges $E_{\circlearrowleft}$ connecting nodes within each layer graph G^{t_i}, and the set of edges $E_{\curvearrowright}$ connecting the same nodes across different layer graphs G^{t_i} and $G^{t_j} (i \neq j)$, where $E = E_{\circlearrowleft} \cup E_{\curvearrowright}$.

The construction of correlation graph consists of three parts. First, because we aim to use existing stores to estimate the numbers of potential customers for query locations, we connect each unlabeled node $v_\circ \in V_\circ$ to all the labeled nodes $u_\bullet \in V_\bullet$ within each time period. Second, owing to the fact that the potential customer number of a store is highly correlated to its historical values, we connect each unlabeled node within time period t_j to the corresponding unlabeled node within each of its previous time period $t_i (i < j)$. Third, since the potential

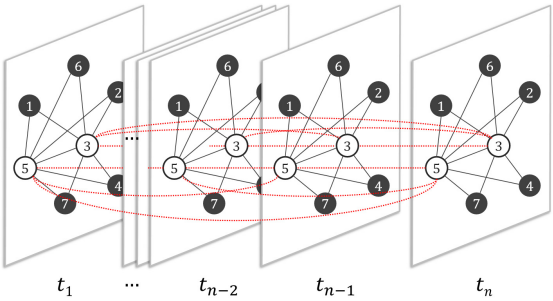

Fig. 3. An illustration to the correlation graph.

customer numbers of near-by locations have higher possibility to be close to one another (due to sharing similar volume of crowds), each unlabeled node $v_\circ \in V_\circ$ is connected to the near-by unlabeled ones $u_\circ \in V_\circ$ within a geographical radius r ($r = 200$ meters), where $dist(u_\circ, v_\circ) < r$.

We illustrate the correlation graph using Figure 3. There are seven locations as nodes $V = \{v_1, v_2, ..., v_7\}$, in which five are labeled nodes $V_\bullet = \{v_1, v_2, v_4, v_6, v_7\}$ and two are unlabeled nodes $V_\circ = \{v_3, v_5\}$. Since there are n time periods $t_1, t_2, ..., t_n$, we construct n layer graphs sharing the same node set V. In each layer graph G^{t_i}, a set of internal edges $E_{\circlearrowleft}$ are constructed, as illustrated using bold lines. For any two layer graphs G^{t_i} and $G^{t_j} (i \neq j)$, we construct a set of external edges $e_{ij} \in E_{\frown}$ connecting the same unlabeled nodes $v_\circ \in V_\circ$ between G^{t_i} and G^{t_j}, as shown using dash lines. We will describe the way to determine edge weights W^{t_i} through *feature-aware location correlation* in the following.

5.3 Location Correlation

Learning edge weights in CG from location features plays a key role in the estimation of potential customer numbers for unlabeled nodes. We aim to model the *correlation* between a labeled and an unlabeled node as their edge weight. The idea is that for a certain time period, if two locations with higher *correlation*, they tend to have *closer* potential customer numbers. In other words, for an unlabeled node $v \in V_\circ$, its number of potential customer will be close to that of the location with higher correlation to each other. The geographical, mobility, and social features are exploited to characterize the *correlation* between locations. In general, two locations whose features have lower difference should share closer potential customer numbers, while higher difference should make their potential customer numbers far away. Nevertheless, various feature might have different degree of effect on the correlation of potential customer numbers between the feature difference. For example, though lower feature differences of *Area Popularity* and *Density* between two locations make their potential customer numbers close, one should have more significant effect on the other. Therefore, the importance of separate feature should be considered.

For a certain store, we estimate the *feature-aware location correlation* based on their differences with respect to each feature. Then we combine the values of feature-aware location correlation of all the features through weighted sum. The weight multiplied by each feature location correlation will be determined based on its predictability of potential customers.

Definition 7: Feature-aware Location Correlation (FLC). Given a particular feature f_k, the *feature-aware location correlation* $flc_{f_k}(u(t_i), v(t_j))$ between nodes u and v, $(u, v) \in E$, in time periods t_i and t_j respectively can be derived from their feature difference $flc_{f_k}(u(t_i), v(t_j)) = \Delta f_k(u(t_i), v(t_j))$, where Δf_k is their feature difference, defined by $\Delta f_k = \|\mathbf{f_k}(u(t_i)) - \mathbf{f_k}(v(t_j))\|$.

Given a set of features $F = \{f_1, f_2, ..., f_m\}$, we combine *feature-aware location correlation* value $flc(u(t_i), v(t_j))$ between nodes u and v, $(u, v) \in E$, in time periods t_i and t_j, via the weighted sum of their correlation flc_{f_k}, given by:

$$flc(u(t_i), v(t_j)) = \exp(-\sum_{k=1}^{m} \pi_k \times flc_{f_k}(u(t_i), v(t_j))), \tag{1}$$

where π_k is the weight of feature f_k. The combined correlation is considered as the edge weight $w_{u(t_i),v(t_j)} = flc(u(t_i), v(t_j)$ between nodes u and v in CG.

Feature-based Top Store Detection. To determine feature weight π_k, we use the values of each feature f_k on existing stores to detecting stores with higher check-in numbers, and if f_k leads to higher precision scores, it will be assigned a higher feature weight. We use *Precision@X%* to evaluate the goodness of each feature. An instance is a store at a certain time period t_i, and there are $|S| \times |T|$ instances in total, where S is the set of all the stores. We denote the set of all instances to be ST. By setting $X\% = 10\%, 20\%, 30\%$ of stores with top/higher check-in numbers, we define the scores of *Precision@X%* as $|ST_{f_k,X\%} \cap ST_{X\%}|/|ST_{X\%}|$, where $ST_{X\%}$ is the set of stores with top $X\%$ check-in numbers, and $ST_{f_k,X\%}$ is the set of stores with top $X\%$ values of feature f_k. Features with higher precision scores provide more benefit on estimating potential customer numbers of stores. Therefore, we compute the weight π_k of each feature f_k by normalizing the average precision scores of f_k over all the features, $\pi_k = [0, 1]$.

5.4 Customer Inference Algorithm

We estimate the potential customer numbers of arbitrary locations over time $t_1, t_2, ..., t_n$ using the correlation graph. The idea is to iteratively update the number of potential customers $pc(v_\circ)$ of each unlabeled node $v_\circ$ until the change of their potential customer numbers converges. Since the correlation of potential customer numbers among locations or stores is described by the correlation graph, we compute the potential customer number $pc(v_\circ)$ from its neighboring labeled or unlabeled nodes. This is fulfilled by averaging the potential customer numbers of $v_\circ$'s neighbors, which are weighted by edge weights. Since the correlation graph provides benefits on modeling the temporal and spatial correlation

Algorithm 1. Potential Customer Estimation (PCE)

Input: (a) a set of existing store locations $V_\bullet$ with existing potential customer numbers $pc(v_\bullet)$ ($v_\bullet \in V_\bullet$), (b) a set of query locations $V_\circ$, (c) the time periods to be observed $T = t_1, t_2, ..., t_n$

Output: $pc(v_\circ(t_i))$, where $v_\circ \in V_\circ$ and $t_i \in T$

1 $V \leftarrow V_\bullet \cup V_\circ$;
2 $\mathbf{f_k}(v) \leftarrow$ extracting feature f_k, $k = 1, 2, ..., m$, $v \in V$;
3 Construct CG from V and $\mathbf{f_k}(v)$, $v \in V$;
4 Compute feature weights π_k by *Precision@X%* with normalization;
5 $w_{uv} \leftarrow \exp(-\sum_{k=1}^{m} \pi_k \times flc_{f_k}(u(t_i), v(t_j)))$;
6 Initialize the potential customer number of each unlabeled node $pc(v_\circ(t_i)) \leftarrow \sum_{u \in N(v_\circ(t_i)) \& u \in V_\bullet} w_{v_\circ(t_i),u} \times pc(u)$;
7 $\Delta avgPc \leftarrow \frac{1}{|V_\circ|} \times \sum_{v_\circ(t_i) \in V_\circ} pc(v_\circ(t_i))$;
8 **while** $\Delta avgPc > \epsilon$ **do**
9 **for** $v_\circ(t_i) \in V_\circ$ **do**
10 $pc(v_\circ(t_i)) \leftarrow \sum_{u \in N(v_\circ(t_i))} w_{v_\circ(t_i),u} \times pc(u)$;
11 $\Delta avgPc \leftarrow \frac{1}{|V_\circ|} \times \sum_{v_\circ(t_i) \in V_\circ} pc(v_\circ(t_i))$;
12 return $pc(v_\circ(t_i))$.

of potential customers, stores with higher correlation with $v_\circ$ contribute more weights on the estimation of potential customer numbers for unlabeled nodes.

We give the complete algorithm of Potential Customer Estimator (PCE) in Algorithm 1. We first use both existing stores (i.e., labled nodes $V_\bullet$) and query locations (i.e., unlabeled nodes $V_\circ$) to construct the correlation graph based on the extracted features f_k for each node (line 1-3). With the *feature-based top store detection*, we can derive the weight of each feature π_k and use feature weight to initialize the edge weight w_{uv} in the correlation graph (line 4-5). Then we can further initialize the potential customer number of each unlabeled node $pc(v_\circ(t_i))$ from the set of $v_\circ(t_i)$'s neighboring labeled nodes $N(v_\circ(t_i)) \subset V_\bullet$ (line 6). We also initialize the difference of the average potential customer numbers between iterative rounds $\Delta avgPc$ by the sum of the initialized potential customer numbers of unlabeled nodes (line 7). In the iterative updating (line 8-11), we continue adjusting the potential customer numbers of unlabeled nodes based on those of its neighboring labeled and unlabeled nodes and the edge weights. This iterative process will terminate until $\Delta avgPc$ converges.

6 Experiments

We conduct experiments to exhibit the performance the proposed PCE model. The objective is three-fold. First, we aim to understand the effectiveness of PCE, comparing to a series of competitors. Second, we are eager to know whether or not PCE can successfully detect the locations with higher potential customer

numbers. Three, we wonder how different combinations of features and different feature settings affect the performance of PCE.

6.1 Evaluation Plans

Competitive Methods. We compare PCE with a series of competitive methods, which are divided into four categories. The first is spatial k-nearest neighbors; the second category is two interpolation-based methods, i.e., Inverse Distance Weighting and Ordinary Kriging; the third is two conventional learning methods (i.e., Artificial Neural Network and Support Vector Regression); and the fourth is two state-of-the-art semi-supervised learning methods, i.e., Co-Training and Radial Basis Function-based SSL. Note that SVR is one of the methods that have the best performance popularity ranking on Geo-Spotting [8].

- **Spatial k-Nearest Neighbors (kNN)** considers the average potential customer number from the potential customer numbers of the k closet geographical neighboring locations as the estimated value.
- **Inverse Distance Weighting (IDW)** is a well-known interpolation method [4]. IDW assigns values of unlabeled locations by calculating the weighted averages of the values available on labeled locations. Locations lower geographical distances have higher weights.
- **Ordinary Kriging (OK)** [17] is a state-of-the-art method of spatial point interpolation. The prediction is calculated as weighted averages of geographical neighbors, in which the weights are determined by finding the *semi-variogram* values for instances between known locations and the semivariogram values for instances between each unknown location and all known locations. Then a set of simultaneous equations are solved by minimizing the estimation error of each unknown location.
- **Artificial Neural Network (ANN)** with the commonly-used back propagation technique is used as another baseline. The constructed ANN contains one hidden layer in the experiments for the generality. We set a linear function for the neurons in the input layer and assign a sigmoid function for those in the hidden and output layers.
- **Support Vector Regression (SVR).** A version of SVM for regression is choose chosen to estimate the potential customer numbers. SVR utilizes the historical check-in data on locations as the training data and learn a cost function to build the predictive model.
- **Co-Training (CT)** is proposed by Nigam and Ghani [15] and serves as the state-of-the-art method for learning the correlation between real values. The co-training model consists of two separated classifiers. One is a spatial classifier based on artificial neural network to model the spatial correlation of labels. The other is a temporal classifier based on a *linear-chain conditional random field* (CRF) [10] to model the temporal dependency of labels.
- **Radial Basis Function-based Semi-supervised Learning (SSL)**, which is a state-of-the-art graph-based learning method [23], serves as a strong competitor. To apply RBF-SSL, the potential customer numbers of query locations within each time period are estimated separately, in which a graph is

constructed for each time period based on geographical distance. In addition, we quantize the potential customer numbers of locations as ten discrete labels, and consider the mean value of the predicted label to be the result.

Evaluation Metrics. We use two metrics in the experiments: *Hit Rate* and *Normalized Discounted Cumulative Gain* (NDCG). For a location $v_\circ$ in the query set of locations $V_\circ$ within time period $t_i \in T(T = t_1, t_2, ..., t_m)$, assume its estimated potential customer number is $pc(v_\circ(t_i))$ and the ground-truth potential customer number is $\tilde{pc}(v_\circ(t_i))$. Then the *hit rate* is defined as:

$$HitRate = \frac{\sum_{v_\circ \in V_\circ, t_i \in T} hit(pc(v_\circ(t_i)), \tilde{pc}(v_\circ(t_i)))}{|V_\circ| \cdot |T|}, \tag{2}$$

where $hit(pc(v_\circ(t_i)), \tilde{pc}(v_\circ(t_i))) = 1$ if $\tilde{pc}(v_\circ(t_i)) - \gamma \leq pc(v_\circ(t_i)) \leq \tilde{pc}(v_\circ(t_i)) + \gamma$, otherwise: $hit(pc(v_\circ(t_i)), \tilde{pc}(v_\circ(t_i))) = 0$, where the parameter γ determines the strictness of the evaluation through varying the *granularity* of the ground-truth potential customer numbers. A higher γ value indicates a loose generous evaluation and every methods would have higher accuracy in general; a lower γ value refers to a strict evaluation, and thus the accuracy tends to be lower for different methods. We choose to have a strict evaluation with $\gamma = 30$. The second evaluation metric is NDCG [7]. We use NDCG to estimate the ranking quality between the potential customer numbers estimated by a method and the ground-truth potential customer number. Higher scores of Hit Rate and NDCG mean better performance.

Basic Settings. To evaluate PCE, we use the potential customer numbers of stores of three retail chains, Starbucks (SB), McDonald's (MC) and Dunkin' Donuts (DD). We choose such three retail chains because their stores are three of the most popular and the most widely scattered in New York City. Such three retail chains are considered as three evaluation subsets. For each retail chain, we divide its stores into training and test parts. Assume there are n_S stores and n_T time periods, we randomly select 80% stores as training instances ($80\% \times n_S \times n_T$) and the other 20% stores are regarded as test instances, whose locations are used as the query and their potential customer numbers within each time period are removed and served as the ground truth. For the parameters used in the experiments, we have the following settings by default: (a) the geographical neighboring radius of feature extraction $r = 200$ meters, (c) two categories of time period granularity are considered: week and month, (d) all the three feature sets, geographical features (GF), temporal mobility features (TMF), cumulative mobility features (CMF), and social features (SF) are used together, and (e) the strictness parameter $\gamma = 30$ for the evaluation metric of accuracy.

Detailed Plans. To reach the three goals mentioned above, we have the following four detailed evaluation plans. The first is the *general evaluation*, in which the proposed PCE is compared to seven competitors. The general evaluation will be conducted under time periods of weeks and months for the three retail chains.

The second is the *top potential customers evaluation*, which is designed to understand whether or not the proposed PCE can successfully detect locations with higher potential customer numbers. The third is the *feature importance evaluation*. Through reporting the performance of PCE using different combinations of feature sets, including GF, TMF, CMF, and SF we can know which feature is more important in the estimation of potential customer numbers. The fourth is the *feature range evaluation*. Recall the feature values computed are constrained to a certain geographical radius r of neighborhood. We aim to present the performance by varying the radius r, to understand the predictability of geographical areas.

6.2 Experimental Results

General Evaluation. The results for the three retail chains under time periods of weeks and months are shown in Table 1. We can find that PCE significantly outperforms all of the competitors under all the cases. We think such promising results come from not only the investigation of location correlation as well as feature-aware location ranking, but also the simultaneous consideration of spatial and temporal dependency between locations and stores in the correlation graph. However, most of the competitors that purely learn the correlation between features and potential customer numbers. In more details, it can be observed that the accuracy is hard to exceed 0.8 under time periods of weeks, especially for the competitors. We think it is because we choose a strict evaluation with $\gamma = 30$ in the experiments. In addition, we can find that the performance of months is much better than that of weeks. It is due to the fact that the popularity value accumulated in each month is higher than that of each week. Therefore, the potential customer numbers of months tend to be a bit far apart from each other and make it a bit easier to be estimated.

Top Potential Customers Evaluation. We test if the proposed PCE is able to detect the locations with higher potential customer numbers. Following the settings described in the section of Feature-based Top Store Detection and using and the same evaluation metric $Precision@X\%$, we aim to present the estimated potential customer numbers by PCE by varying the percentage of locations with the highest potential customer numbers from 5% to 35%. We report the $Precision@X\%$ scores in Figure 4. We can find that the precision scores by PCE can have 0.8 precision scores for top 20% stores with the highest potential customer numbers. Such results exhibit the practical usages of PCE on estimating and finding hot zones in a city, and demonstrate the effectiveness of using stores as sensors to estimate the numbers of potential customers.

Feature Importance Evaluation. To understand which feature set is more important on potential customer estimation, we report the performance of different combinations of feature sets (i.e., GF, TMF, CMF, and SF) using PCE, as shown in Table 2. We can find that comparing to GF, CMF, and SF, TMF obtains the better results with higher scores of NDCG scores in general under

Table 1. General Evaluation Results on weeks and months.

	Week						Month					
	nDCG			HitRate			nDCG			HitRate		
	SB	MC	DD	SB	MC	DD	SB	MC	DD	SB	MC	DD
kNN	0.15	0.30	0.25	0.11	0.12	0.12	0.26	0.33	0.39	0.13	0.14	0.14
IDW	0.17	0.30	0.25	0.11	0.12	0.12	0.24	0.31	0.38	0.13	0.14	0.14
OK	0.18	0.35	0.28	0.19	0.24	0.23	0.29	0.34	0.39	0.16	0.16	0.15
ANN	0.53	0.58	0.60	0.52	0.54	0.53	0.57	0.61	0.64	0.69	0.67	0.69
SVR	0.58	0.61	0.62	0.58	0.60	0.56	0.62	0.64	0.65	0.72	0.73	0.75
CT	0.56	0.67	0.65	0.56	0.52	0.60	0.64	0.70	0.69	0.74	0.74	0.75
SSL	0.63	0.71	0.69	0.63	0.66	0.68	0.68	0.74	0.72	0.74	0.74	0.74
PCE	**0.71**	**0.79**	**0.78**	**0.79**	**0.84**	**0.81**	**0.76**	**0.82**	**0.80**	**0.83**	**0.88**	**0.88**

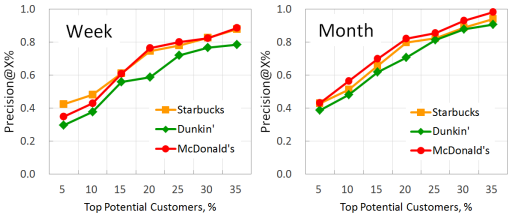

Fig. 4. Evaluation of Top Potential Customers using PCE, by varying the percentage of stores with the highest potential customer numbers.

both weeks and months and for all the three retail chains. We think the reason could be the TMF is capable of describe the neighboring human flows at the separate time periods while CMF can only capture the historical volume of human flow traveling in the neighborhood of a location, which reflects the total numbers of potential customers. As for GF and SF, what it captures is the properties and distributions of location categories and social activities in the neighborhood, and thus cannot directly exhibit the volume of potential customers. Therefore, GF and SF derives the worse estimation accuracy than CMF and TMF.

Feature Range Evaluation. Since the features extracted are constrained within a certain neighborhood via the radius r (in meters), we would like to which r is more effective and leads to better performance in PCE. The results are shown in Figure 5. We can find the performance of using $r = 300$ is the best for time periods of weeks while using $r = 200$ is the best for months. Too small or large radius values r leads to worse performance because features extracted constrain on a small area could not fully and precisely describe the neighborhood while constraining on a large area might include irrelevant features. The feature neighborhood radius $r = 300$ and $r = 200$ also is quite close to and responses to the optimal neighborhood radius 200 meters suggested by urban planning [14].

Table 2. Feature Importance Evaluation: the NDCG scores of different feature sets.

	Geographical Feat. (GF)						Temporal Mobility Feat. (TMF)					
	Week			Month			Week			Month		
	SB	DD	MC	SB	DD	MC	SB	DD	MC	SB	DD	MC
kNN	0.159	0.256	0.303	0.256	0.385	0.331	0.159	0.256	0.303	0.256	0.385	0.331
IDW	0.171	0.252	0.305	0.244	0.377	0.312	0.171	0.252	0.305	0.244	0.377	0.312
OK	0.189	0.284	0.352	0.289	0.393	0.342	0.189	0.284	0.352	0.289	0.393	0.342
SVR	0.322	0.332	0.455	0.355	0.514	0.648	0.541	0.582	0.656	0.589	0.635	0.731
ANN	0.329	0.342	0.452	0.368	0.522	0.651	0.517	0.568	0.632	0.552	0.622	0.722
CT	0.334	0.351	0.452	0.369	0.524	0.652	0.533	0.578	0.649	0.561	0.642	0.729
SSL	0.341	0.382	0.464	0.431	0.529	0.668	0.557	0.580	0.658	0.562	0.657	0.712
PCE	**0.349**	**0.401**	**0.481**	**0.462**	**0.552**	**0.681**	**0.582**	**0.619**	**0.713**	**0.663**	**0.685**	**0.756**
	Cumulative Mobility Feat. (CMF)						**Social Feat. (SF)**					
	Week			Month			Week			Month		
	SB	DD	MC	SB	DD	MC	SB	DD	MC	SB	DD	MC
kNN	0.159	0.256	0.303	0.256	0.385	0.331	0.132	0.247	0.280	0.194	0.375	0.316
IDW	0.171	0.252	0.305	0.244	0.377	0.312	0.168	0.255	0.271	0.195	0.379	0.284
OK	0.189	0.284	0.352	0.289	0.393	0.342	0.172	0.259	0.337	0.206	0.380	0.309
SVR	0.426	0.368	0.531	0.526	0.588	0.645	0.393	0.327	0.498	0.475	0.561	0.583
ANN	0.418	0.354	0.520	0.513	0.543	0.638	0.391	0.333	0.460	0.428	0.517	0.604
CT	0.426	0.391	0.536	0.535	0.596	0.659	0.432	0.369	0.555	0.396	0.520	0.637
SSL	0.435	0.408	0.519	0.529	0.585	0.673	0.471	0.388	0.526	0.466	0.594	0.650
PCE	**0.587**	**0.455**	**0.603**	**0.613**	**0.633**	**0.724**	**0.477**	**0.412**	**0.539**	**0.498**	**0.604**	**0.704**

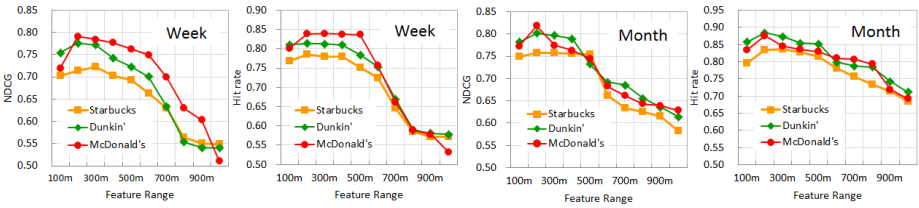

Fig. 5. Feature Range Evaluation, by varying the neighrborhood radius r using PCE.

7 Conclusion

Being able to acquire the knowledge about where and when the customers will show up can lead to many useful applications, including determining the locations of new business, choosing the right time and place to host campaign to maximize the advertise effect. This paper proposes a method to estimate the number of potential customers in an urban area. We leverage stores as a kind of sensors to estimate the potential customers of of any location during any given time span. A *PCE* model is developed and validated with promising performance. In the future, we aim to go beyond location-based services and further consider more heterogeneous urban information into the modeling of potential customers, such as traffic status, weather, and near-by activities.

References

1. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment (2008)
2. Chen, Z., Liu, Y., Wong, R.C.-W., Xiong, J., Mai, G., Long, C.: Efficient algorithms for optimal location queries in road networks. In: ACM SIGMOD (2014)
3. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: ACM KDD (2011)
4. Donald, S.: A two-dimensional interpolation function for irregularly-spaced data. In: ACM National Conference (1968)
5. Fu, Y., Ge, Y., Zheng, Y., Yao, Z., Liu, Y., Xiong, H., Yuan, N.J.: Sparse real estate ranking with online user reviews and offline moving behaviors. In: IEEE ICDM (2014)
6. Hsieh, H.-P., Lin, S.-D., Zheng, Y.: Inferring air quality for station location recommendation based on urban big data. In: ACM KDD (2015)
7. Jarvelin, K., Kekalainen, J.: Cumulated gain-based evaluation of IR techniques. ACM TOIS (2002)
8. Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., Mascolo, C.: Geo-spotting: mining online location-based services for optimal retail store placement. In: ACM KDD (2013)
9. Kisilevich, S., Mansmann, F., Keim, D.: P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geotagged photos. In: COM.Geo (2010)
10. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: ICML (2001)
11. Li, Y., Steiner, M., Wang, L., Zhang, Z.-L., Bao, J.: Exploring venue popularity in four-square. In: IEEE INFOCOM (2013)
12. Lima, A., Musolesi, M.: Spatial dissemination metrics for location-based social networks. In: ACM UbiComp (2012)
13. Liu, Y., Wei, W., Sun, A., Miao, C.: Exploiting geographical neighborhood characteristics for location recommendation. In: ACM CIKM (2014)
14. Mehaffy, M., Porta, S., Rofe, Y., Salingaros, N.: Urban nuclei and the geometry of streets: The emergent neighborhoods' model. Urban Design International (2010)
15. Nigam, K., Ghani, R.: Analyzing the effectiveness and applicability of co-training. In: ACM CIKM (2000)
16. Monreale, A., Pinelli, F., Trasarti, R., Giannotti, F.: Where next: a location predictor on trajectory pattern mining. In: ACM KDD (2009)
17. Oliver, M.A., Webster, R.: Kriging: a method of interpolation for geographical information systems. IJGIS (1990)
18. Sadilek, A., Kautz, H., Bigham, J.P.: Finding your friends and following them to where you are. In: ACM WSDM (2012)
19. Scellato, S., Mascolo, C., Musolesi, M., Latora, V.: Distance matters: geo-social metrics for online social networks. In: WOSN (2010)
20. Tiwari, S., Kaushik, S.: User category based estimation of location popularity using the road GPS trajectory databases. Geoinformatica (2014)
21. Ying, J.-C., Lee, W.-C., Weng, T.-C., Tseng, V.S.: Semantic trajectory mining for location prediction. In: ACM SIGSPATIAL GIS (2011)
22. Zhang, C., Shou, L., Chen, K., Chen, G., Bei, Y.: Evaluating geo-social influence in location-based social networks. In: ACM CIKM (2012)
23. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML (2003)