Location Recommendations for New Businesses using Check-in Data

Bahaeddin Eravci[1], Neslihan Bulut[2], Cagri Etemoglu[3] and Hakan Ferhatosmanoglu[4]

[1, 2, 4]Dept. of Computer Engineering, Bilkent University, Ankara, Turkey
[3]Turk Telekom Research Labs, Istanbul, Turkey
E-mail: [1]beravci@gmail.com, [2]neslihan.bulut@gmail.com, [3]cagri.etemoglu@turktelekom.com.tr, [4]hakan@cs.bilkent.edu.tr

***Abstract*—Location based social networks (LBSN) and mobile applications generate data useful for location oriented business decisions. Companies can get insights about mobility patterns of potential customers and their daily habits on shopping, dining, etc. to enhance customer satisfaction and increase profitability. We introduce a new problem of identifying neighborhoods with a potential of success in a line of business. After partitioning the city into neighborhoods, based on geographical and social distances, we use the similarities of the neighborhoods to identify specific neighborhoods as candidates for investment for a new business opportunity. We present two solutions for this new problem: i) a probabilistic approach based on Bayesian inference for location selection along with a voting based approximation, and ii) an adaptation of collaborative filtering using the similarity of neighborhoods based on co-existence of related venues and check-in patterns. We use Foursquare user check-in and venue location data to evaluate the performance of the proposed approach. Our experiments show promising results for identifying new opportunities and supporting business decisions using increasingly available check-in data sets.**

***Index Terms*—location based social networks, business decision systems, spatio-temporal data mining**

I. Introduction

Increasingly more users share their personal data on Internet such as status updates and opinions. This phenomenon has increased the number of content generators to billions and created a diverse community of people resembling the actual population day-by-day diverging from a biased pool of technology enthusiasts. GPS (Global Positioning System) and other positioning systems (wifi, GSM network) have added the geographical location dimension to further enhance the user experience and pave the way for new location based applications. Many people continuously share their location, Mostly through mobile apps, which are usually called check-ins. A variety of LBSNs have flourished in this niche market, such as Foursquare/Swarm, Facebook places, Tinder, Sports Tracker, Everymove, Zombies!run!, Yelp, Groupun, Untappd. We seeks ways to utilize this data effectively for business intelligence and decisions.

In this paper, we propose a new data analytics problem and a solution to identify existence of business opportunities in a city neighborhood using location based data like Foursquare check-ins. We focus on the case where an investor wishes to start a new venue in the city on a particular line of business. Finding the best location for a venue is a classical problem with a variety of solutions including recent data analytics approaches such as bichromatic reverse nearest neighbor (RNN) queries [1]. Given a set of locations L and a set of customers C, bichromatic RNN queries return the customers from set C for which the queried location is the nearest neighbor. Bichromatic RNN queries can be used to infer the best locations for a venue to attract the highest number of customers[2]. These methods usually work on a single type of business, such as opening a new branch for a fast-food chain or a wireless service provider. Gathering data about potential customers was difficult in traditional terms but it is now much easier thanks to abundant information provided by location based applications such as check-in data sets. We also see general methods to extract patterns as subsets of features co-located using the concept of proximity [3].

Check-ins can help to identify implicit relationships between venues through correlations of their customers, e.g. people who shop at a specific supermarket tend to also visit coffee shops. We hypothesize that check-ins for similar categories, even when the venues are in different regions, can be used to predict the potential visits to a new venue. We verify this intuition and use similarity of businesses based on their check-in patterns and similarity of a category in different regions to identify neighborhoods where there is a high probability of existence for a given venue type is relatively high. We investigate two methods, one based on Bayesian inference and correlation of categories and another on a collaborative filtering based on similarity of neighborhoods. Our solution analyzes the categories of businesses and the commonalities of neighborhoods to recommend a region in which the user can open a new venue. Following the same co-location premise one can also identify business categories that are missing or have a high potential for a given neighborhood and recommend these categories as new business opportunities.

We first generate neighborhoods of a city through clustering based on a combination of social and spatial distances. We then follow two approaches for recommendation, each with its own strength and trade-off, in solving the problem. The first one is a probabilistic neighborhood selection (PNS) where Bayesian inference is used to calculate the posterior probability of the specific line of business based on the inputs of all the other business types in the same region. This method takes into account both the prior probability of the specific line of business and the evidence (i.e., the existence of the other business in the neighborhood) to make the recommendation. We also develop an efficient approximation (PNS-A) which is a voting algorithm on the "related categories" of the business line. Related categories are the different businesses which tend to be co-located with the category of interest. We analyze the correlation in different neighborhoods using the training data to identify the set of related categories for each line of business. We show that this analysis is an approximation of PNS method. Our second approach is an adaptation of the concept of "collaborative filtering" to this new problem. We propose collaborative neighborhood filtering (CNF) that finds a set of similar neighborhoods with respect to the queried region, and recommend business categories that are common in this set, yet have low (or no) representation in the queried region. This aims to decide whether a particular line of business has a potential in the area of interest by looking at similar other regions.

We have performed experiments on location based social network (LBSN) data of New York from Foursquare to validate and compare the different methods for this new problem. The experiments first focus on the different variability of the social distance when finding the neighborhoods. We have compared the recommendation with the ground truth of whether that specific line of business indeed exists in the recommended neighborhoods to assess the performance of the solutions. The experiments also investigate how the number of neighborhoods found in the city affects the accuracy of the recommendation.

In summary, the main contributions of this paper are as follows:

- To the best of our knowledge, this is the first framework which leverages LBSN data for new venue recommendations without any domain specific user intervention.
- We propose a probabilistic neighborhood selection algorithm to identify suitable regions by maximizing the posterior probability for the given business type by taking into account prior probabilities and the existing business types. We also propose a majority voting method on "related business categories" which we show to be an effective approximation of the Bayesian posterior probability.
- We present a new application of the concept of collaborative filtering for this new problem.
- The experiments have shown encouraging results and also fortified our hypothesis that LBSN data can be useful for making new business recommendations.

The outline of the paper is as follows. In Section 2, we first formally define the problem and present observations from real data that support the intuition that certain types of businesses are co-located. We then present the main framework and sub-components of the framework: finding the neighborhoods in the city, two perspectives and the solutions for the recommendation respectively in Section 2. We present the dataset, experimental setup and the performance results of the proposed algorithms in Section 3. Section 4 concludes the paper.

II. Method

In this section, we define the problem, present the proposed setting, and outline our solutions. The common definitions include the following: U_i references a user i, $\mathbf{U}$ references a set of users, V_i references a business venue i, $\mathbf{V}$ references a set of venues, C_j references a category or line of business, $\mathbf{C}$ references a set of categories. We have C_j values for each V_i based on the LBSN data and category of a particular V_i is shortly denoted as $V_i.cat$. We use bold fonts for sets and $|\ |$ denotes the cardinality of the given set.

A. Problem Definition

We first define region, $\mathcal{N}_i$ (neighborhood i) as a connected area in the city which includes different types and number of business venues (can also be expressed as livehood, geographical region, trade area, etc.). $\mathcal{N}_i$s can be found in different ways including a pure geographic perspective or a combination of activity and locality features such as the livehood concept by Cranshaw et al [4].

In our first problem, we seek to recommend the user a set $\mathcal{N}_i$s ($\boldsymbol{\mathcal{N}_{rec}}$) for which a venue in a specific business category C_j is estimated to be a "successful" business decision. We assume that existence of a venue with the line of business searched translates to a "successful" business case in the recommended region. We define this query in the proposed framework as follows:

Query : Given a specific category of business C_j, recommend a set of neighborhoods, $\boldsymbol{\mathcal{N}_{rec}}$, for which existence of a business venue is highly probable ($\boldsymbol{\mathcal{N}_{rec}} \subset \boldsymbol{\mathcal{N}}$). $|\boldsymbol{\mathcal{N}_{rec}}|$ will also be specified such that the query asks for a specific number (n) of neighborhood areas.

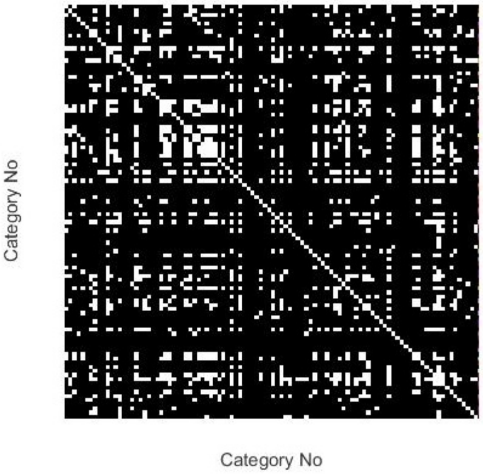

Figure 1. Correlation between categories (cosine similarity (between clusters in the city which have the related two categories) higher than 0.5)

B. Observations from Real Data

We observe from daily life that some groups of business venues tend to cluster together in different parts of the city, e.g., coffee shops are co-located near restaurants with a high probability. Figure 1 illustrates this observation using real LBSN data as a correlation matrix for different categories (white points depict pairwise categories which are co-located throughout the city). Our approach utilizes this correlation information in recommending new business venues by analyzing the present venues and identifying missing or less represented ones which could have a high business potential.

To enable such an analysis, we first cluster the business venues across the city based on their similarities of locations and/or the sets of users visiting the venues to define the neighborhoods. We then develop methods to decide whether the business venue of category C_j can be successful in a specific neighborhood $\mathcal{N}_i$. Algorithm 1 shows the main flow of the analysis and captures both of the queries presented previously.

Algorithm 1 High-level algorithm for business recommendation system

```
$C_j$, specific category of business is given
$LBSN$, location based social network data is given
$k$, number of neighborhoods
$n$, number of recommended neighborhoods ($|\boldsymbol{\mathcal{N}_{rec}}|$)
$\boldsymbol{\mathcal{N}} = FindNeighborhoods(k)$, partition the city
$\boldsymbol{\mathcal{N}_{rec}} = BusinessRecommend(LBSN, \boldsymbol{\mathcal{N}}, C_j, n)$
```

C. Finding Neighborhoods

We utilize the venues and their check-ins to first partition the city into neighborhoods, as proposed in [4]. Equation 1 defines the distance between V_i and V_j as a weighted sum of their geographical ($GDist$) and social distance ($SDist$) with a tuning parameter α.

$$D(V_i, V_j) = \alpha\ GDist(V_i, V_j) + (1 - \alpha)\ SDist(V_i, V_j) \quad (1)$$

$$GDist(V_i, V_j) = [(V_i.lat - V_j.lat)^2 + (V_i.long - V_j.long)^2]^{(0.5)}$$
$$SDist = Jaccard(\text{Users of } V_i, \text{Users of } V_j)$$

For simplicity, we use the flat Earth model and approximate the distance as the Euclidean distance between coordinates of the venues. It is approximately linear proportional to the geodesic distance if the distance is small with respect to the radius of the sphere which is the case in our application. Social distance on the other hand is the Jaccard distance between the users of the venues which signifies the common users visiting the respective venues.

After the definition of the distance which incorporates the properties of neighborhood one can use any clustering scheme to partition the data. For our experiments, we have used k-means with k (number of clusters) and α (distance weight parameter). Figure 2 depicts an example case using the check-in data of New York.

D. Probabilistic Neighborhood Selection (PNS)

Section II-B has explained with evidence from the data that some business types are co-located in the neighborhoods, i.e., if one exists the other exists as well. The statement, C_j is highly probable if C_m exists in a neighborhood, means that $P(C_j = 1|C_m = 1)$ is high. Query 1 can be defined as finding the posterior probability defined in Equation 2 for all $C_j \in \mathbf{C}$.

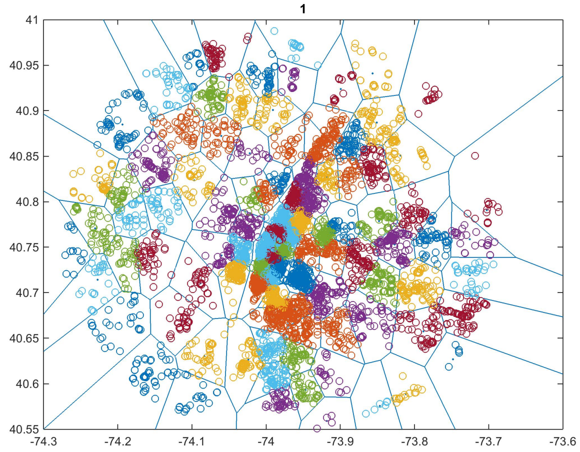

Figure 2. Neighborhood structure for $k = 100$ and $\alpha = 1$ (each color represents a neighborhood with the respective Voronoi cell)

$$P(C_j = 1|C_1, C_2, \ldots, C_{j-1}, C_{j+1}, \ldots, C_J) \quad (2)$$

where $C_m = \begin{cases} 1, & \text{if a venue of category } C_m \text{ exists in } \mathcal{N}_i \\ 0, & \text{otherwise} \end{cases}$

We note that the posterior probability is a conditional probability of all categories except the very category we are looking for in the given neighborhood $\mathcal{N}_i$. C_m values can also be considered as continuous variables but since the data is not sufficient for healthy estimates of the probabilities we opted to use binary existence variable.

Using Bayes theorem we are able rewrite the posterior as following:

$$P(C_j = 1|C'_j) = \frac{P(C'_j|C_j = 1)\ P(C_j = 1)}{\sum_{l=0}^{1} P(C'_j|C_j = l)\ P(C_j = l)} \quad (3)$$

where $C'_j = C_1, C_2, \ldots, C_{j-1}, C_{j+1}, \ldots, C_J$.

Prior probability ($P(C_j = 1)$) in the expression is the probability of C_j existing in any neighborhood without any knowledge of the variation of the venues in the specific neighborhood. Likelihood, $P(C'_j|C_j = 1)$, is the parameter that we have to learn from the data which incorporates the relation of the different class of venue with the venue type C_j. The denominator is the normalization parameter to find the correct posterior probability.

We also make the assumption, that every category is independent of each other when the category C_j is considered.

$$P(C_l, C_m|C_j) = P(C_l|C_j).P(C_m|C_j) \quad \forall l, m \neq j$$

Using this assumption we can further manipulate Equation 3 into:

$$P(C_j = 1|C'_j) = \frac{P(C_j = 1) \prod_{C_m \in C'_j} P(C_m|C_j = 1)}{\sum_{l=0}^{1} P(C_j = l)\ \prod_{C_m \in C'_j} P(C_m|C_j = l)}$$

We calculate the posterior probability for each $\mathcal{N}_i$ to find the most probable neighborhood. The overall algorithm is outlined in Algorithm 2.

Algorithm 2 Algorithm for Probabilistic Neighborhood Selection

```
$C_j$, $LBSN$, $\boldsymbol{\mathcal{N}}$, $n$ are given
Calculate all prior probabilities $P(C_m|C_j = 1)$ and $P(C_m|C_j =)$ for $C_m \in C'_j$ using known data ($LBSN$)
for $\forall \mathcal{N}_i \in \boldsymbol{\mathcal{N}}$ do
    $Posterior(i) = 0$, posterior probability of $C_j$ in $\mathcal{N}_i$
    Calculate $Posterior(i) = P(C_j = 1|C'_j)$ using priors and $\mathcal{N}_i$ data
end for
$\boldsymbol{\mathcal{N}_{rec}} = \mathcal{N}_i$s with highest $Posterior$ where $|\boldsymbol{\mathcal{N}_{rec}}| = n$
```

E. Approximation of PNS (PNS-A) using Related Category Analysis

In this approach (PNS-A), we first find the correlations of the business categories and form a set of related categories for each category C_j. Our assumption here is that we can simplify the model if related categories exist in the particular $\mathcal{N}_i$ for the recommendation. The details of the probabilistic interpretation and its relation with PNS is given in the Appendix.

The method analyzes neighborhoods by checking the existence of venues from each business category and forms the binary $\mathcal{N}\_CAT$ matrix which is defined as in Equation 4. Related categories are defined according to the column similarities of $\mathcal{N}\_CAT$ matrix. A threshold value is also applied over the pairwise similarities of the columns.

$$\mathcal{N}\_CAT\{i,j\} = \begin{cases} 1, & \text{if venue of category } C_j \text{ exists in } \mathcal{N}_i \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The related categories information can be used to recommend a new business for any neighborhood if the number of related categories in the region is correlated with the success of the business venue.

We find the correlations of the different categories, and form a set of related categories, $\mathbf{RelCat}$, according to the correlations for each category C_j. Items in $\mathbf{RelCat}$ are identified as the categories whose co-location with venues of type C_j is most probable.

The method starts by analyzing the city structure by looking at the existence of venues of specific categories in the different neighborhoods and forms the binary $\mathcal{N}\_CAT$ matrix defined as in Equation 4. $\mathcal{N}\_CAT$ matrix records the neighborhood-business type information. The method proceeds with further analysis on the $\mathcal{N}\_CAT$ matrix.

We find the correlation between categories using the similarities between columns of $\mathcal{N}\_CAT$ matrix. $\mathbf{RelCat}$ set is found by thresholding the pairwise Jaccard similarities.

After finding a set of related categories for each category we can use this set to recommend a category for any new neighborhood. The analysis of the venues of the new neighborhood with respect to the related categories determines the estimated probability of presence of business venues. We expect that the number of related categories in the region is proportional with the potential success of the business venue. The details of the method with the overall outline of the proposed system is given in Algorithm 3.

Algorithm 3 Algorithm for approximation of PNS (PNS-A)

```
$C_j$, $LBSN$, $MinRelCat$, $\boldsymbol{\mathcal{N}}$, $n$ are given
$\mathcal{N}\_CAT$ matrix is calculated per Eq. 4
$\mathbf{RelCat}$ = FINDRELCAT($C_j$,$\mathcal{N}\_CAT$)
Recommended set of neighborhoods, $\boldsymbol{\mathcal{N}_{rec}} = \emptyset$
for $\forall \mathcal{N}_i \in \boldsymbol{\mathcal{N}}$ do
    $NoOfRelCat(i)$ = number of related categories in $\mathcal{N}_i$
end for
$\boldsymbol{\mathcal{N}_{rec}} = \mathcal{N}_i$s with highest $NoOfRelCat$s where $|\boldsymbol{\mathcal{N}_{rec}}| = n$
```

F. Collaborative Neighborhood Filtering (CNF)

We address the same problem in a perspective to be solved with collaborative filtering approaches. The main idea in collaborative filtering is to find similar entities (neighborhoods in our case) to the queried one and identify commonalities in these entities as a recommendation. In our context, we find similar neighborhoods to the given $\mathcal{N}_i$ and make use of the $\mathcal{N}\_CAT$ matrix defined Equation 4 which consists of existence information of each business category C_j in each neighborhood $\mathcal{N}_i$. This forms the neighborhood-business category matrix which is analogous to user-item matrix in traditional recommender systems.

We pose the similarity problem using the $\mathcal{N}\_CAT$ matrix and use Jaccard index for the similarity calculations which is used especially in binary cases in align with our problem. We exclude C_j when calculating the similarity since we are querying about this particular business type. Jaccard index calculation for our problem is provided in Equation 5.

$$\begin{aligned} J(\mathcal{N}_i, \mathcal{N}_m, C_j) &= \frac{\mathcal{N}_i \cap \mathcal{N}_m}{\mathcal{N}_i \cup \mathcal{N}_m} \\ &= \frac{\sum_{n \in C'_j} min\{\mathcal{N}\_CAT(i,n), \mathcal{N}\_CAT(m,n)\}}{\sum_{n \in C'_j} max\{\mathcal{N}\_CAT(i,n), \mathcal{N}\_CAT(m,n)\}} \\ &\text{where } C'_j = (1, 2, \ldots, j-1, j+1, \ldots, |\mathbf{C}|-1, |\mathbf{C}|) \end{aligned} \tag{5}$$

For a given $\mathcal{N}_i$, we retrieve a particular set of similar neighborhoods, $\boldsymbol{Sim\mathcal{N}_i}$ as the basic model for the particular $\mathcal{N}_i$. The size of this set ($|\boldsymbol{Sim\mathcal{N}_i}|$) is usually chosen as a small fraction of the whole dataset (denoted as F).

After finding the similar set, we estimate the likelihood of C_j in $\mathcal{N}_i$ by analyzing the patterns of C_j in the neighborhoods of $\boldsymbol{Sim\mathcal{N}_i}$. We calculate the likelihood as the weighted sum of the evidence in the data. The weights we use are the similarity index that we have calculated in the previous parts.

Likelihood of C_j in $\mathcal{N}_i$ is calculated using the Equation 6 with respect of $\boldsymbol{Sim\mathcal{N}_i}$.

$$L(\mathcal{N}_i, C_j) = \frac{\sum_{\mathcal{N}_m \in \boldsymbol{Sim\mathcal{N}_i}} J(\mathcal{N}_i, \mathcal{N}_m, C_j).GA\_MAT(m,j)}{\sum_{\mathcal{N}_m \in \boldsymbol{Sim\mathcal{N}_i}} J(\mathcal{N}_i, \mathcal{N}_m, C_j)} \tag{6}$$

We run the procedure for each $\mathcal{N}_i$ and calculate the respective $L(\mathcal{N}_i, C_j)$ to make the recommendation decision. The overall algorithm is given in

Algorithm 4 Algorithm for business recommendation using collaborative neighborhood filtering

```
$C_j$, $LBSN$, $\boldsymbol{\mathcal{N}}$, $n$, $F$ are given
$\mathcal{N}\_CAT$ matrix is calculated per Eq. 4
Recommended set of neighborhoods, $\boldsymbol{\mathcal{N}_{rec}} = \emptyset$
for $\forall \mathcal{N}_i \in \boldsymbol{\mathcal{N}}$ do
    $L(i) = 0$, likelihood of $C_j$ in $\mathcal{N}_i$
    Find $\boldsymbol{Sim\mathcal{N}_i} = NearestNeighbor(\mathcal{N}\_CAT, \mathcal{N}_i, C_j, F)$ using Equation 5
    Calculate $L(i)$ using Equation 6
end for
$\boldsymbol{\mathcal{N}_{rec}} = \mathcal{N}_i$s with highest $L(i)$ where $|\boldsymbol{\mathcal{N}_{rec}}| = n$
```

III. Performance Evaluation

We have performed experiments on real data collected from Foursquare to validate the introduced framework and evaluated the accuracy of the proposed methods.

A. Dataset and Experimental setup

The dataset used includes check-in data for New York city collected from Foursquare from 12 April 2012 to 16 February 2013 with venue location and user check-in information [5]. After removing the venues with less than 5 check-ins, the data set has 179,468 checkins and 9,986 venues with a high density around the Manhattan area which is expected.

Dataset has a total of 251 different category of business venues. Since our aim is to propose new business opportunities, we have selected a subset of the categories that fits into our application. We have included categories like "bar" and "restaurants" and excluded venues like "zoo", etc. which are irrelevant to our cause. We have excluded cases where the business venue is very rare (e.g. "Afghan Restaurant") which we have little information about its correlation with different line of businesses. These types of business are also considered irrelevant since the probability of opening many new business venues will also be very low.

We partition the dataset into training and test sets for fair evaluation of PNS and PNS-A. Training dataset is used to calculate the related parameters ($P(C_m|C_j = 1)$ and $P(C_j = l)$ in Equation 4 and to find the related categories using Algorithm 3) used in the respective methods. The test and the training sets are selected with respect to latitude of the cluster center which enables a near bisection of the dataset. Partitioning of the training and test sets was performed by selecting the clusters whose centroid point is greater than the mid-point, $40.75°$ latitude for this dataset. We do not use any such partitioning in the collaborative filtering case and use the whole dataset in our analysis since it does not need partitioning.

We have calculated two different performance measures. One is the accuracy based on top-n retrieval. In this case we feed the system with a C_j and the test neighborhoods and retrieve n neighborhoods ($\boldsymbol{\mathcal{N}_{rec}}$) which the system recommends. We have used four different values, $n = 1, 3, 5, 10$, for our testing purposes which encompasses the practical scenarios where an investor is not expected to request more than 10 neighborhoods for investment recommendation. Accuracy is defined as follows:

$$Accuracy = \frac{1}{|\mathbf{C_{tested}}|} \sum_{C_j \in \mathbf{C_{tested}}} \frac{1}{|\boldsymbol{\mathcal{N}_{rec}}|} \sum_{\mathcal{N}_i \in \boldsymbol{\mathcal{N}_{rec}}} 1_{\mathcal{N}_i}(C_j)$$

where indicator function $1_{\mathcal{N}_i}(C_j) =$

$$\begin{cases} 1, & \text{if a venue of category } C_j \text{ exists in } \mathcal{N}_i \\ 0, & \text{otherwise} \end{cases}$$

Another measure we use for performance evaluation is the area under curve of a precision-recall graph. The system's output can be thresholded for a final decision based on $MinRelCat$ in related categories method and $MinPosterior$ in the Bayesian inference method. Using these parameters, we can control the system's precision and recall performance. We have experimented by varying $MinRelCat$ in the range of $[0 : |\mathbf{RelCat}|]$ with one increments and the $MinPosterior$ in the range $[0, 1]$ with 0.1 increments. These experiments have provided results on precision and recall levels which we plotted in a precision-recall graph.

We have defined "Baseline" as the probability of a particular business line in geographical area calculated using the training data (ratio of the number of regions which includes at least one venue of type of interest and the total number of regions).

Based on the precision-recall curves we have calculated area under curve (AUC) to capture the information within the precision curves. The average value of the AUC for different $\mathcal{N}_i$s in the test data is considered as the performance indicator. This performance measure is given only for related categories and Bayesian inference methods since the collaborative filtering based approach is not suitable for precision-recall analysis.

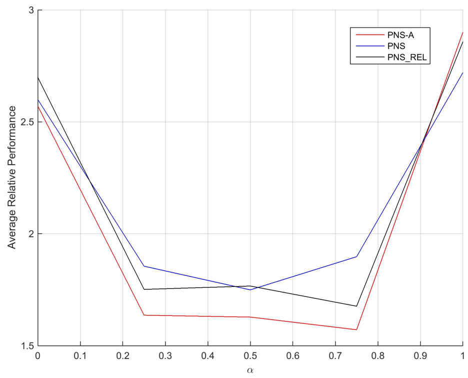

Figure 3. Relative performance of methods

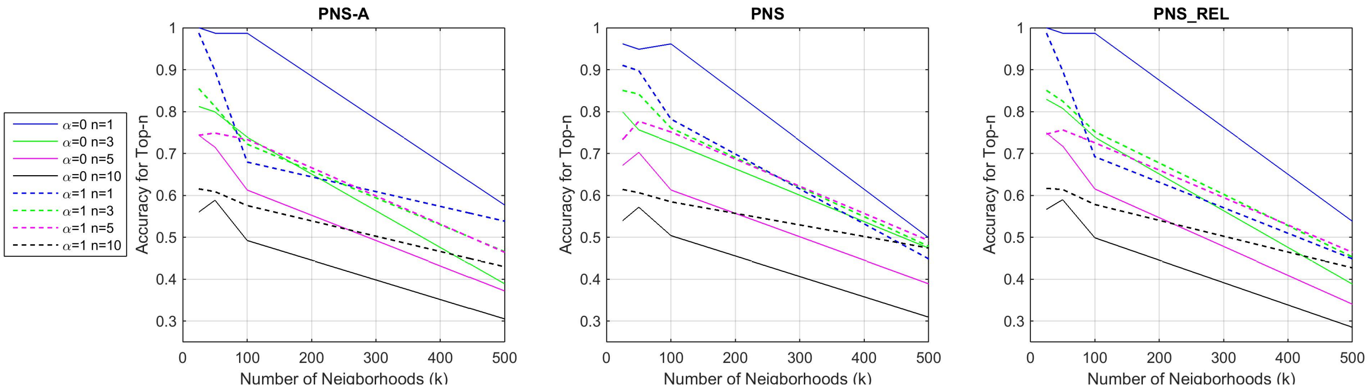

Figure 4. Top-n performance of methods

B. Experimental Results and Discussions

1) Neighborhood Analysis: We have performed clustering experiments with $\alpha \in [0, 0.25, 0.5, 0.75, 1]$ where $\alpha = 0$ is social distance only, $\alpha = 1$ is geographical distance only and the others in between. We have also experimented on number of clusters by varying it as $k \in [25, 50, 100, 500]$.

We have provided the clusters and their respective voronoi diagrams for the case of $k = 100$ and $\alpha = 1$ in Figure 2 as an illustration of the neighborhood structure. We observe that if we increase the affect of social, the neighborhood structure does change and the clustering structure does not conform with its respective voronoi cell. Some points are related with cluster centroids but are not in the respective voronoi cell of the centroid which is the case in $\alpha = 1$ case. For the $\alpha = 0$ case, the clusters are completely out of sync with neighborhoods which prevents the definition of continuous geographical region. The neighborhood radius are inversely proportional with the number of clusters (k) which is expected.

2) Performance of the methods: In this section we present the performance results of the solutions: Probabilistic neighborhood selection (PNS), approximate probabilistic neighborhood selection (PNS-A), Bayesian inference using the related categories only (PNS_REL), Collaborative neighborhood filtering (CNF), Collaborative neighborhood filtering using the related categories only (CNF_REL).

We first present the accuracy results of the first three methods starting with the area under curve (AUC) values for different α and k values. The performance decreases as the number of neighborhood increases. This is mainly because of the fact that the prior probability (probability of C_j being in any $\mathcal{N}_i$) decreases with k.

AUC for PNS-A method is superior than the other methods. To correctly assess the performance difference between the methods, we have calculated the relative performance increase with respect to a baseline which is defined as the ratio of method's performance and the baseline probability for that specific C_j. We have averaged the ratio over different k and different C_j values to find the average relative performance. The relative performance with respect to α is depicted in Figure 3. The results show that the system best performs with $\alpha = 1$ and $\alpha = 0$ cases which are the social-only and location-only clustering methods. We also observe that the location-only clustering achieves higher accuracies than the social-only clustering for all the methods. Even though the PNS-A method is simpler in terms of both computation and mathematical complexity, it achieves more accurate results than the other methods.

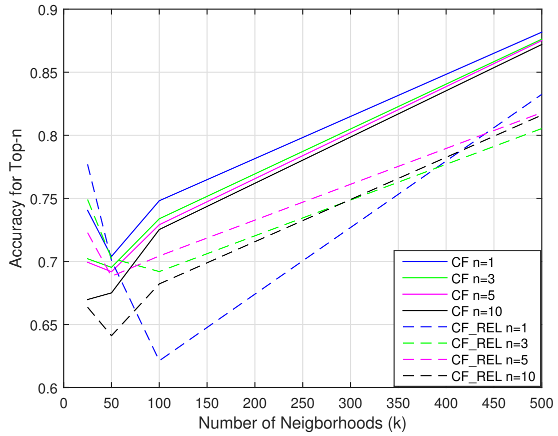

Figure 5. Top-n performance of collaborative filtering methods

Table I
AVERAGE RESULTS FOR DIFFERENT METHODS AND PARAMETERS

		$n = 1$	$n = 3$	$n = 5$	$n = 10$
PNS-A	$\alpha = 0$	**0.8878**	0.6848	0.6109	0.4865
	$\alpha = 1$	0.7756	0.7137	0.6724	0.5574
PNS	$\alpha = 0$	0.8429	0.6891	0.5942	0.4814
	$\alpha = 1$	0.7596	**0.7329**	**0.6885**	**0.5699**
PNS_REL	$\alpha = 0$	0.8782	0.6912	0.6058	0.4853
	$\alpha = 1$	0.7564	0.7201	0.6731	0.5590

We now present the accuracy of the top-n recommendation results of the methods with two different perspectives. The results for the first class of methods which uses global features of the data (PNS,PNS-A,PNS_REL) are presented in Figure 4. We have seen a similar effect in terms of α and have only presented $\alpha = 1$ and $\alpha = 0$ cases which are the two competing case. We observe a similar pattern where the accuracy declines with the number of neighborhoods. We present the average accuracy values for different methods and parameters in Table I in which the highest accuracy values for each n value is given in bold. We see that for $n = 1$ case $\alpha = 0$ and related categories method is the best pair for recommender performance. In the other cases of n, we observe that $\alpha = 1$ case is superior and all the methods perform similarly with slight differences.

Top-n accuracy results for collaborative filtering methods exhibit an interestingly different pattern as shown in Figure 5. We have plotted the $\alpha = 1$ case since we have observed that this case has been superior to all the other α cases in each of these methods. One of the interesting behavior of these methods is clear that collaborative filtering methods are superior in cases where we have high number of neighborhoods. This was the case where the previous methods have relatively failed. Especially in the $k = 500$, CNF methods perform with accuracy more than 0.85 where the previous methods had accuracy below 0.6 which is a considerable difference. We also observe that the performance gap for different n values decrease with the increase of k. We base this observation due to the fact that as k increases we are able to find more similar neighborhoods in the city. We also observe that collaborative filtering without the related categories modification achieves more accurate results. We also have to note that CNF does not need any training process but needs distance calculations for nearest neighbor calculations for each query.

3) Performance for Different Business Categories: In this section we discuss the variation of performances for different lines of businesses and give more qualitative results. To clarify the analysis we have chosen a case where $n = 3$, $k = 100$. We look at two methods: related categories and collaborative filtering which both have accuracy around 0.75. We also preferred the geographical-only clustering ($\alpha = 1$) since this parameter choice has better results for our case. The results are shown in Figure 6. We observe clearly a binary structure in the related category case where the method either performs very high or very low (0). There seems to be a positive correlation between prior probability and the accuracy of the related categories method. This is mainly caused from the fact that if there is not enough data in the training set for these methods the recommendation performs poorly and vice versa. From these results we can see that a hybrid system can be used to increase the accuracy further. We also see that the system can recommend with very high accuracy for typical business categories such as american restaurant, bakery, bank, bar, coffee shop, deli/bodega, fast food restaurant, etc.

IV. Conclusion

We have proposed a business recommendation framework based on analyzing similarities of geographic neighborhoods using check-in data sets. Our approach identifies a new neighborhood in which a specific type of business venue is expected to be present. The result can be used to identify a promising neighborhood for a new venue. We have proposed two main solutions: one on Bayesian inference and its approximation using a majority voting scheme over related categories (based on correlations of business categories), and another on collaborative filtering. We have shown with experiments on real data that the proposed solution can recommend with accuracy 2-3 times better than a baseline approach.

Check-in data sets can be utilized in other creative ways for new business and investment opportunities. We plan to work on a more refined recommendation system where the system can estimate not just the existence of a particular business line but also “how successful” the business will be by looking at the expected quantitative values of the check-in data of the venues. While this extension is rather straightforward, it needs a larger data-set to correctly estimate the distributions of these continuous variables.

Acknowledgment

This study was funded in part by Turk Telekom. Bulut was supported in part by The Scientific Technological Research Council of Turkey (TUBITAK) under TUBITAK 2232 grant no: 114C124.

References

[1] F. Korn and S. Muthukrishnan, “Influence sets based on reverse nearest neighbor queries,” in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD ’00. New York, NY, USA: ACM, 2000, pp. 201–212. [Online]. Available: http://doi.acm.org/10.1145/342009.335415

[2] J. Huang, Z. Wen, J. Qi, R. Zhang, J. Chen, and Z. He, “Top-k most influential locations selection,” in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’11. New York, NY, USA: ACM, 2011, pp. 2377–2380. [Online]. Available: http://doi.acm.org/10.1145/2063576.2063971

[3] Y. Huang, S. Shekhar, and H. Xiong, “Discovering colocation patterns from spatial data sets: A general approach,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 16, no. 12, pp. 1472–1485, Dec. 2004.

[4] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh, “The livehoods project: Utilizing social media to understand the dynamics of a city,” *ICWSM’12*, 2012.

[5] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, “Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 1, pp. 129–142, Jan 2015.

Appendix

We explain the mathematical intuitions for the related category method. First, we illustrate what high jaccard distance means in terms of probability. Jaccard index is defined as follows:

$$\begin{aligned} J(C_j, C_m) =& \frac{C_j \cap C_m}{C_j \cup C_m} \\ =& \frac{|C_j^1 \cap C_m^1|}{|C_j^1 \cap C_m^1| + |C_j^0 \cap C_m^1| + |C_j^1 \cap C_m^0|} \end{aligned}$$

where C_j^1 is the set of $\mathcal{N}$s where C_j exists and C_j^0 otherwise.

If $J(C_j, C_m)$ is high (ideally equal to 1) we can neglect the terms $|C_j^0 \cap C_m^1| + |C_j^1 \cap C_m^0|$, which really are the discrepancies where in some particular $\mathcal{N}$ one of the business class exists

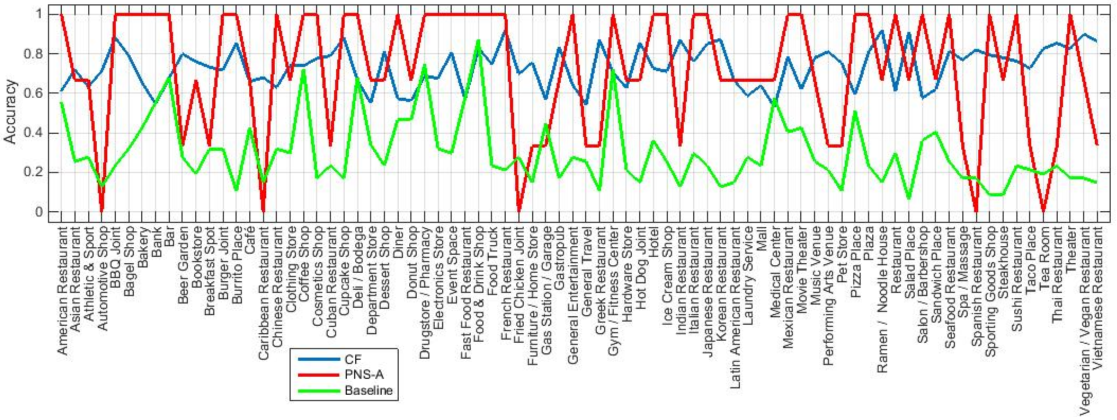

Figure 6. Accuracy of different methods with respect to different business types

and the other does not exist. This observation leads to the following approximated conditional probabilities:

$$P(C_j = 1|C_m = 1) = \frac{|C_j^1 \cap C_m^1|}{|C_m^1|} = \frac{|C_j^1 \cap C_m^1|}{|C_j^1 \cap C_m^1| + |C_j^0 \cap C_m^1|} \simeq 1 \tag{7}$$

$$P(C_j = 0|C_m = 0) = \frac{|C_j^0 \cap C_m^0|}{|C_m^0|} = \frac{|C_j^0 \cap C_j^0|}{|C_j^1 \cap C_m^0| + |C_j^0 \cap C_m^0|} \simeq 1 \tag{8}$$

We now approximate Equation 3 i.e. $P(C_j = 1|C'_j)$. We have illustrated the Venn diagram for the event spaces in our case in Figure 7. We were not able to show all the intersections for the clarity of the figure. The black shaded areas are where the $\mathcal{N}$s are densely populated and using the approximations of Equation 7 and 8 we assume that we do not have any events except the black regions for those sets. For the unrelated categories, we assume that they are evenly distributed (shaded gray) in C_j^1 and C_j^0 sets. C_r^1s are C_r^1s are events of existence and non-existence of a particular category of business respectively. C_r and C_{r+1} is a related category for C_j where C_u and C_{u+1} are the opposite.

For calculation of the probability of $C_j = 1$ in $\mathcal{N}_i$, we make the following definitions; C_R^1 and C_R^0 are the sets of C_is which are present and absent respectively in $\mathcal{N}_i$ and are related to C_j. C_U^1 and C_U^0 are opposite sets which are not related (Unrelated) to C_n. $C'_j = C_R^1 \cup C_R^0 \cup C_U^1 \cup C_U^0$ and since all these cases are mutually exclusive we can also say that $|C'_j| = |C_R^1| + |C_R^0| + |C_U^1| + |C_U^0|$.

$$\begin{aligned}
P(C_j = 1|C'_j) &= \frac{|C_j^1 \cap C'_j|}{|C'_j|} = \frac{|C_j^1 \cap (C_R^1 \cup C_R^0 \cup C_U^1 \cup C_U^0)|}{|C'_j|} \\
&\simeq \frac{|C_j^1 \cap C_R^1| + |C_j^1 \cap C_R^0| + |C_j^1 \cap (C_U^1 \cup C_u^0)|}{|C'_j|} \\
&\simeq \frac{|C_j^1 \cap C_R^1| + |C_j^1 \cap (C_U^1 \cup C_U^0)|}{|C'_j|}
\end{aligned}$$

Assuming number of $\mathcal{N}_i$s in each $|C_j^1 \cap C_r^1| = |C_j^0 \cap C_r^0| = n \; \forall r \in R$. and $|C_j^1 \cap (C_U^1 \cup C_U^0)| = |C_j^0 \cap (C_U^1 \cup C_U^0)| = m$ we can further manipulate the equation:

$$\begin{aligned}
P(C_j = 1|C'_j) &\simeq \frac{|C_j^1 \cap C_R^1| + |C_j^1 \cap (C_U^1 \cup C_U^0)|}{|C_R^1| + |C_R^0| + |C_U^1 \cup C_U^0|} \\
&\simeq \frac{|C_j^1 \cap C_R^1| + |C_j^1 \cap (C_U^1 \cup C_U^0)|}{|C_R^1| + |C_R^0| + |C_j^1 \cap (C_U^1 \cup C_U^0)| + |C_j^0 \cap (C_U^1 \cup C_U^0)|} \\
&\simeq \frac{n.|C_R^1| + m}{n.|C_R^1| + n.|C_R^0| + m + m} \quad \text{(assuming } n.|C_R^1| >> m) \\
&\simeq \frac{|C_R^1|}{|C_R^1| + |C_R^0|}
\end{aligned} \tag{9}$$

Equation 9 shows that the posterior probability can be approximated as the ratio of number of existent related categories to the total number of related categories which is used in our related categories method.

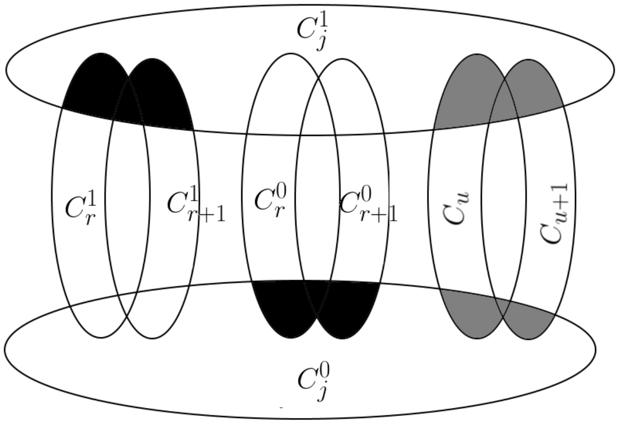

Figure 7. Events in the probability space