Where to Place Your Next Restaurant? Optimal Restaurant Placement via Leveraging User-Generated Reviews

Feng Wang
Department of Computer Science
Hong Kong Baptist University
Hong Kong, China
fwang@comp.hkbu.edu.hk

Li Chen
Department of Computer Science
Hong Kong Baptist University
Hong Kong, China
lichen@comp.hkbu.edu.hk

Weike Pan
College of Computer Science and Software Engineering
Shenzhen University
Shenzhen, China
panweike@szu.edu.cn

ABSTRACT

When opening a new restaurant, *geographical placement* is of prime importance in determining whether it will thrive. Although some methods have been developed to assess the attractiveness of *candidate* locations for a restaurant, the accuracy is limited as they mainly rely on traditional data sources, such as demographic studies or consumer surveys. With the advent of abundant user-generated restaurant reviews, there is a potential to leverage these reviews to gain some insights into users' preferences for restaurants. In this paper, we particularly take advantage of user-generated reviews to construct predictive features for assessing the attractiveness of candidate locations to expand a restaurant. Specifically, we investigate three types of features: review-based *market attractiveness*, review-based *market competitiveness* and *geographic characteristics* of a location under consideration for a prospective restaurant. We devise the three sets of features and incorporate them into a regression model to predict the number of check-ins that a prospective restaurant at a candidate location would be likely to attract. We then conduct an experiment with real-world restaurant data, which demonstrates the predictive power of features we constructed in this paper. Moreover, our experimental results suggest that market attractiveness and market competitiveness features mined solely from user-generated restaurant reviews are more predictive than geographic features.

CCS Concepts

•Information systems → Data mining;

Keywords

Optimal restaurant placement; User-generated reviews; Market attractiveness features; Market competitiveness features; Geographic features

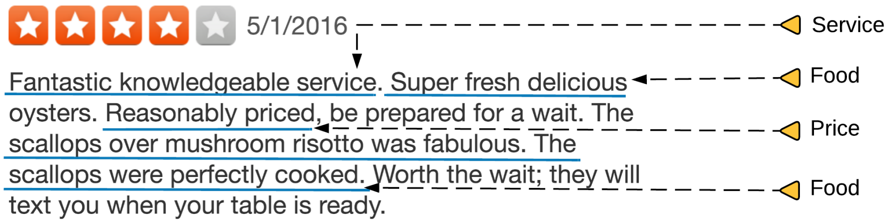
CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA
© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2983323.2983696

1. INTRODUCTION

When a restaurant thrives in one geographic location (location refers to a general area within a city), the temptation can be great to expand it to new locations. Each location specifically has its own advantages and disadvantages with regard to different types of restaurants [10]. For example, access, visibility, population demographics, traffic patterns, and the presence of complementary businesses including other restaurants have been recognised as major factors affecting a restaurant's success or failure [9]. Thus, business success in one location does not necessarily guarantee success in another location.

In related work, the restaurant location selection mainly relies on data obtained from demographic studies, consumer surveys and the like [4, 16], which, however, are usually very expensive to acquire. Recently, some studies have attempted to automatically assess location quality based on the analysis of spatial distribution [12], but they require access to datasets that are not usually publicly available. Hence, the issue of how to identify the right location for a restaurant in an accurate and timely fashion, particularly by using publicly accessible data, remains a challenging problem that requires extensive study.

In this work, we are engaged in the problem of selecting an optimal new place to expand an existing restaurant. The priority when we are planing the expansion of a restaurant to a new location is to profile the restaurant's existing *customers*' preference. When we understand the preference of customers a restaurant attracts, we may then identify proper locations that have high concentrations of users with similar preferences (i.e., potential customers). The primary novelty of our work lies in leveraging a particular form of user-generated content, restaurant reviews, which are publicly available in local directory services such as *Yelp.com.* Usually, the customers often write reviews to express detailed opinions about multi-faceted aspects of a restaurant, as illustrated in Figure 1. The descriptive reviews of restaurants contributed by customers provide a valuable opportunity to understand why a restaurant in a particular location is attractive or not. We can hence discover the cultural idiosyncrasies of that area and gain insight into the preferences of restaurant customers. Specifically, given a collection of restaurant reviews, we can apply aspect-based opinion mining techniques [14] to extract the aspects of a restaurant (e.g., food, price and service), as well as inferring sentiment polarity for each aspect. With the results of aspect-based opinion mining, we can first infer how often


5/1/2016

Fantastic knowledgeable service. Super fresh delicious oysters. Reasonably priced, be prepared for a wait. The scallops over mushroom risotto was fabulous. The scallops were perfectly cooked. Worth the wait; they will text you when your table is ready.

Figure 1: An example of a restaurant review that expresses *positive* opinions about the restaurant's food, price and service.

an aspect is commented in reviews of a restaurant. If its customers often commented an aspect, it implies that the customers have high interest in that aspect. Hence, such aspect distribution can be used to profile the preference of customers for a restaurant. Meanwhile, we can summarize the sentiment polarities of aspects into sentiment scores for that restaurant, which can be used to represent how satisfied the customers are with that restaurant.

As the major contribution of this paper, we investigate two types of geographic dependency features between locations and restaurants, which are mined solely from restaurant reviews: 1) *market attractiveness* features which measure whether a restaurant would be attractive to customers who may visit its located place; 2) *market competitiveness* features which measure whether a restaurant would offer competitive quality in comparison with competitors located in the same place. Besides the above two types of review-based features, we also consider the *geographic* features which can be mined from publicly accessible data source to encode the spatial properties (e.g., the number of restaurants) of a location. In our work, we mainly focus on the case of expanding a chain restaurant to a new location. For individual who wants to identify a good location for a new restaurant, the locations of other similar and successful restaurants can provide important guidelines.

2. OPTIMAL RESTAURANT PLACEMENT

In this section, we formalize the optimal restaurant placement problem as a dyadic predictive task for predicting the number of *check-ins* a prospective restaurant will receive if it is placed in a candidate location.

2.1 Problem Statement

Formally, there is a set of candidate locations $\mathcal{L}$ from which a location $\ell \in \mathcal{L}$ can be chosen for the next place a restaurant is expanded to. As previously noted, our goal is to identify the *optimal* restaurant location ℓ^* that will potentially attract the largest number of visits. Location ℓ can be formally defined by its longitude and latitude coordinates and radius r, as illustrated in Figure 2. Unless otherwise specified, in our experiment, r is set to 200 metres, as that radius is estimated to be the optimal size of a neighbourhood [13]. Our problem can be defined as a dyadic prediction problem in which the dyad is a restaurant-location pair, and concretely described by a set of features.

To solve this problem, we seek to exploit the mined features from reviews to predict the number of visits. The valuable information contained in reviews provides a unique opportunity to arrive at a more holistic understanding of customers, restaurants and locations. For a prospective restaurant, the location that obtains the highest predicted number of visits will be identified as the optimal location. As the starting time varies across restaurants, it is better to use the average number of check-ins *per* month rather than the total number of check-ins to assess the popularity of a restaurant.

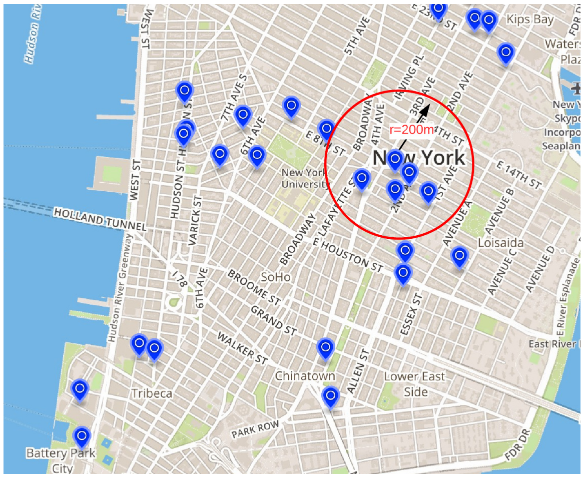


Figure 2: The demonstration of a location, which refers to an area with a radius of $r = 200m$ around a particular point in New York.

2.2 Features

In this section, we first introduce the two types of features: *market attractiveness* features and *market competitiveness* features, which are referred as *review-based* features, as they are mined solely from the review texts and expected to reveal the attractiveness and competitiveness of a location for a specific restaurant. Then, we introduce the third type of features, *geographic* features, to describe the spatial properties of a candidate location.

2.2.1 Review-based Market Attractiveness Features (MAF)

We attempt to construct *market attractiveness* features to measure whether a restaurant would be attractive to customers who may visit its located place. Specifically, to estimate the *market attractiveness* of a restaurant in a new location, the preference of the restaurant's target customers could be a valuable source of information. As one of the highlights in our work, we are engaged to infer how often each aspect is commented in the restaurant's reviews to represent its customers' preference. With this goal, we adopt a LDA-based topic model [5] to identify aspects[1] from restaurant reviews, by following the assumption used in [18], which states the words in one sentence of a review can be referred to the same aspect.

[1] For the sake of simplicity, the terms "topic" and "aspect" are exchangeable through this paper, unless otherwise specified.

Formally, to profile the customers' preference of a given restaurant, we treat the reviews associated with a restaurant $v \in V$ as a single document d_v (where V is the set of restaurants included in the dataset). Following the idea of LDA model, let $t_{vk} = p(k|v)$ be the probability of document d_v belonging to aspect k, and $\beta_{kw} = p(w|k)$ be the probability that aspect k generates word w, then the likelihood of observing the whole review document corpus is defined as:

$$L(\mathbf{W}|\mathbf{\Theta}, \boldsymbol{\beta}) = \prod_{v} \prod_{w} (\sum_{k=1}^{K} t_{vk}\beta_{kw})^{n(v,w)} \tag{1}$$

where $\mathbf{W}$ is the document-word count matrix, $n(v, w)$ is the occurrence count of word w in document d_v, and $\mathbf{\Theta} = \{\boldsymbol{\theta}_v\}_{v=1}^{V}$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_k\}_{k=1}^{K}$ are the parameters to be estimated by maximizing the likelihood function $L(\mathbf{W}|\mathbf{\Theta}, \boldsymbol{\beta})$. Then, the preference of the restaurant v's customers can be represented as the aspect distribution in document d_v, such as $\boldsymbol{\theta}_v = \{t_{v1}, \dots, t_{vK} | 0 \le t_{vk} \le 1 \text{ and } \sum_{k=1}^{K} t_{vk} = 1\}$.

Furthermore, for a location $\ell \in \mathcal{L}$, its customers' preference $\boldsymbol{\theta}_\ell \in \mathbb{R}^K$ can be taken as the aggregation of all restaurants in location l. Concretely, each aspect probability $t_{\ell k}$ ($k \in [1, K]$) can be defined as

$$t_{\ell k} = \frac{1}{Z} \sum_{v \in A(\ell,r)} \log(N(v) + 1) t_{vk} \tag{2}$$

where $N(v)$ is the number of reviews of restaurant v, $v \in A(\ell, r)$ indicates that restaurant v is located in location ℓ with radius r, and $Z = \sum_{v \in A(\ell,r)} \log(N(v) + 1)$ is the normaliser ensuring that $\boldsymbol{\theta}_\ell$ lies on a $(K - 1)$ dimensional simplex. Note that this aggregation formula gives higher weight to restaurants with more reviews.

Once the preferences of customers from restaurant v and ℓ are respectively represented as aspect distributions as mentioned above, we define two attractiveness features to reveal the attractiveness of a location for a specific restaurant, which are described as follows.

- *Affinity* attractiveness feature:

  $$x_{v\ell} = \{t_{vk} \times t_{\ell k}, 1 \le k \le K\} \tag{3}$$

  where each entity $t_{vk} \times t_{\ell k}$ is used to indicate the preference affinity between customers of restaurant v and those of location ℓ.

- *Complementary* attractiveness feature:

  $$x_{v\ell} = \{t_{vk} \times (1 - t_{\ell k}), 1 \le k \le K\} \tag{4}$$

  where each entity $t_{vk} \times (1 - t_{\ell k})$ indicates the preference complementary between customers of restaurant v and those of location ℓ.

In particular, as chain restaurants share a brand name (e.g., McDonald's), and often have common menu items, services and advertising, it is reasonable to merge the reviews of different restaurants belonging to the same chain into a single document. Doing so allows preference of customers from restaurants belonging to the same chain to be represented by the same aspect distribution.

2.2.2 Review-based Market Competitiveness Feature (MCF)

In this section, we attempt to construct the *market competitiveness* features to measure whether a restaurant could offer competitive quality relative to restaurants (i.e., competitors) located in the same place. To estimate the *market competitiveness* of a prospective restaurant if it will be located in a new location, we attempt to summarize the sentiment scores over multiple aspects based on opinions expressed in its reviews, to reveal the restaurant's quality.

For this purpose, we apply SentiWordNet [6] to assign each word w with a triple of polarity scores (i.e., positivity $s_+(w)$, negativity $s_-(w)$ and objectivity $s_o(w)$, each in range [0,1], and $s_+(w)+s_-(w)+s_o(w) = 1$). Specifically, for review r, the sentiment score of an aspect k is defined as the average positivity score of all related words:

$$s_{rk} = \frac{\sum_{w \in W(r,k)} s_+(w)}{|W(r, k)|} \tag{5}$$

where $W(r, k)$ denotes the set of words which is related to aspect k included in review r as identified by topic LDA model. Then, for restaurant v, the sentiment score of aspect k is defined as

$$s_{vk} = \frac{\sum_{r \in R(v,k)} s_{rk}}{|R(v, k)|} \tag{6}$$

where $R(v, k)$ denotes the set of reviews of restaurant v, which all commented on aspect k.

Finally, the restaurant v's quality can be represented as the sentiment scores over aspects, such as $\mathbf{s}_v = \{s_{v1}, \dots, s_{vK} | 0 \le s_{vk} \le 1\}$. The market competitiveness feature of a restaurant if it is located in a candidate location is accordingly defined as:

$$x_{v\ell} = \{rank_k(v, \ell), 1 \le k \le K\} \tag{7}$$

where $rank_k(v, \ell)$ represents the rank value of restaurant v if it is ranked together with all existing restaurants within location ℓ in the decreasing order of sentiment scores on aspect k.

2.2.3 Geographic Features (GeoF)

To represent the geographic characteristics of a location, we also consider three specific geographic features inspired by [11]. In our work, we extract these features from the venue information via the public API of Foursquare.com, in which each venue is associated with a geographic coordinate and categories (e.g., *sub way*, *school* and *residential*).

Density refers to the number of restaurants $A(\ell, r)$ located in location ℓ with radius r:

$$x_{v\ell} = |\{v \in A(\ell, r)\}| \tag{8}$$

Intuitively, a dense area implies a greater likelihood of opportunistic visits to the restaurants located there.

Neighbourhood Entropy is an entropy measure of the frequency of restaurant categories, assessing the heterogeneity of restaurants located in location ℓ [11], which can be calculated by:

$$x_{v\ell} = -\sum_{\gamma \in \Gamma} \frac{N_\gamma(\ell, r)}{N(\ell, r)} \times \log \frac{N_\gamma(\ell, r)}{N(\ell, r)} \tag{9}$$

where Γ is the set of all restaurant categories in the entire dataset, $N_\gamma(\ell, r)$ denotes the number of restaurants of category γ in location ℓ with radius r, and $N(\ell, r)$ is the total number of restaurants in location ℓ with radius r. A location with higher entropy value is expected to be more diverse in terms of restaurant categories, whereas low entropy implies that the location is biased towards some specific types of restaurant.

Competitiveness measures the proportion of nearby restaurants with the same type of restaurant that is under consideration with respect to the total number of restaurants located in location ℓ:

$$x_{v\ell} = -\frac{N_{\gamma_v}(\ell, r)}{N(\ell, r)} \tag{10}$$

where γ_v denotes the type of restaurant v. In general, the area with large competitiveness score implies many of restaurants with the same type (i.e., competitors) would share the potential customers.

Jensen Quality measures the spatial interactions between restaurants and venues. In contrast to [11], we treat the Jensen Quality of each venue category as a feature value, rather than simply aggregating Jensen Quality values over all venue categories as a single feature value. Formally, given a restaurant v, for each venue category $\hat{\gamma} \in \hat{\Gamma}$ (where $\hat{\Gamma}$ is the set of all venue categories), the Jensen Quality of venue category $\hat{\gamma}$ for restaurant v is defined as:

$$x_{v\ell}(\hat{\gamma} \to v) = \frac{\sum_{\gamma \in \Gamma(v)} \log(\kappa_{\hat{\gamma} \to \gamma}) \times (N_{\hat{\gamma}}(\ell, r) - \overline{N_{\hat{\gamma}}(\ell, r)})}{|\Gamma(v)|} \tag{11}$$

where $\Gamma(v)$ is the set of categories to which restaurant v belongs, $N_{\hat{\gamma}}(\ell, r)$ denotes the number of venues of category $\hat{\gamma} \in \hat{\Gamma}$ in location ℓ with radius r, $\overline{N_{\hat{\gamma}}(\ell, r)}$ denotes how many venues of category $\hat{\gamma}$ are observed on average around the restaurant of category γ_v, and $\kappa_{\hat{\gamma} \to \gamma_v}$ denotes the inter-type attractiveness coefficients, which are defined as

$$\kappa_{\hat{\gamma} \to \gamma} = \frac{N - N_{\hat{\gamma}}}{N_{\hat{\gamma}} \times N_{\gamma}} \sum_{\ell \in \mathcal{L}} \frac{N_{\gamma}(\ell, r)}{N(\ell, r) - N_{\hat{\gamma}}(\ell, r)} \tag{12}$$

where N, N_{γ} and $N_{\hat{\gamma}}$ denote the number of venues, the number of restaurants of category γ_v and the number of venues of type $\hat{\gamma}$ respectively.

In summary, *density* and *neighbourhood entropy* geographic features are restaurant-independent, but *competitiveness* and *Jensen Quality* features depend on the location and the restaurant's categories. It is obvious that these geographic features can reflect only the general area properties of the locations without considering the restaurant's detailed properties, e.g., food, service and so on.

2.3 Methodology

We formulate the problem of predicting the visit number of a given restaurant as a regression problem. Given a set of input features $\mathbf{x}_{v\ell}$ and a target variable $y_{v\ell}$ (the number of check-ins received *per* month), we exploit different regression methods to predict the target variable. In order to make the variable better fit the assumptions underlying regression As the distribution of target variable $y_{v\ell}$ is strongly positively skewed, we use the logarithm of $y_{v\ell}$ to make the variable better fit the assumptions underling regression[2]. In our experiment, we exploit three different regression algorithms: Ridge regression, support vector regression (SVR) [17], and Gradient Boosted Regression Trees (GBRT) [7], in order to minimize the error between actual and predicted target variables.

To open a new restaurant $\hat{v}$, we first construct the features $\mathbf{x}_{\hat{v}\ell}$ (as defined in Section 2.2) across different candidate locations $\ell \in \mathcal{L}$. Subsequently, we rank the candidate locations based on the predicted values resulting from the trained regression model by applying the constructed features.

3. EXPERIMENT

3.1 Dataset Description

We collected a set of restaurant reviews in New York City across 260 different categories that were posted on Yelp.com from October 2012 to June 2014. In the experiment, we only consider restaurants that received at least 50 reviews to ensure that the features extracted from the reviews are reliable. As shown in Table 1, the resulting restaurant dataset comprises 1,094,717 reviews to 5,220 restaurants. As noted in the introduction, we emphasize selecting the optimal location for a new branch of a restaurant. Therefore, we use chain restaurants in the dataset as the testing set that contains 93 chain brands and a total of 389 restaurants. The other restaurants are treated as the training set.

The number of check-ins, geographic features (as described in Section 2.2.3) and venue information in New York were obtained from the public API of Foursquare.com. The total number of venues is 147,307, covering 588 different categories (see Table 1).

	Statistics
Yelp restaurant dataset	
# Restaurants	5,220
# Reviews	1,094,717
# Avg. reviews	209.72
# Restaurant categories	260
Foursquare venue dataset	
# Venues	147,307
# Venue categories	588

Table 1: Statistical summary of the *restaurant* dataset from Yelp.com and the *venue* dataset from Foursquare.com

To obtain the candidate locations $\mathcal{L}$, we use a density-based spatial clustering method (OPTICS) [1] to segment a city into cells, each of which is treated as a candidate location. As a result, in our experiment, the New York city is segmented into 995 cells.

3.2 Testing Procedure & Evaluation Metrics

Given a set of candidate locations $\mathcal{L}$, our goal is to select the optimal location ℓ^* for a prospective restaurant v. For each candidate location $\ell \in \mathcal{L}$, we first construct features $\mathbf{x}_{v\ell}$ using the method described in Section 2.2. They are then fed into the trained regression model to predict the number of check-ins that restaurant v will receive during a month.

Firstly, we use Rooted Mean Square Error ($RMSE$) to test whether our model can predict the number of visits (i.e., check-ins) precisely:

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{(v,\ell) \in T} (\hat{y}_{v\ell} - y_{v\ell})^2} \tag{13}$$

in which T is the testing dataset containing pairs of chain restaurant and its location, and $\hat{y}_{v\ell}$ and $y_{v\ell}$ are the predicted and ground-truth numbers of check-ins restaurant v receives in location ℓ respectively.

Secondly, we use the Spearman's rank correlation coefficient (ρ) [15] to measure how well the predicted number of check-ins can preserve the relative order of ground-truth

[2] For simplicity, in the following, the logarithm of target variable $y_{v\ell}$ is also denoted as $y_{v\ell}$ unless otherwise specified.

check-ins, defined as:

$$\rho = \frac{1}{|B|}\sum_{b\in B}\left(1-\frac{6\sum_{(v,\ell)\in T_b}\left(r(\hat{y}_{v\ell})-r(y_{v\ell})\right)^2}{|T_b|\left(|T_b|^2-1\right)}\right) \qquad (14)$$

where B represents the set of restaurant brands, T_b represents the set of restaurants belonging to brand b, and $r(\hat{y}_{v\ell})$ and $r(y_{v\ell})$ represent the rank variables based on predicted and ground-truth check-ins respectively.

Thirdly, we treat each chain restaurant brand as a query, and the locations where the chain restaurants are placed as the relevant locations, in order to test whether these relevant locations could be ranked on the top according to the predicted number of check-ins. In this evaluation, we use Mean Average Precision to evaluate the model's ranking accuracy of relevant locations:

$$MAP = \frac{1}{|Q|}\sum_{q\in Q}\frac{\sum_{n=1}^{|\mathcal{L}|}(P@n * rel(n))}{\#\text{ relevant locations for query q}} \qquad (15)$$

where Q is the set of queries (i.e., chain restaurant brands) and $P@n$ indicates the precision at n:

$$P@n = \frac{\#\text{ relevant locations in the top } n \text{ results}}{n} \qquad (16)$$

and $rel(n) = 1$ if the n-th location is relevant, otherwise 0.

3.3 Results and Discussion

3.3.1 *Impact of Parameter K*

In our work, the review-based features (i.e., market attractiveness and market competitiveness features) are extracted from LDA topic model. Hence, the choice of aspect number, K, will affect the prediction accuracy. To determine the optimal number of aspects for constructing the review-based features, we tune parameter K in the range of $[5, 25]$ with a tuning step of 5.

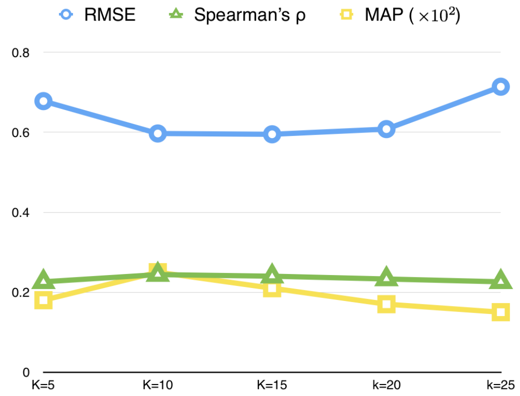


Figure 3: Performance comparison with the review-based features in GBRT algorithm for different aspect numbers K.

Figure 3 shows the changes in terms of $RMSE$, Spearman's ρ and MAP for GBRT algorithm when the aspect number K varies. We can see that the $RMSE$ score reaches the lowest point when $K = 10$, then increases with an increasing number of aspects K. We can also observe that the highest Spearman's ρ (0.244) and MAP (0.0025) are achieved when $K = 10$. For the other two prediction algorithms (i.e., Ridge regression and SVR), the best performance is also achieved when $K = 10$. Hence, we select $K = 10$ as the optimal number of aspects in our experiment.

3.3.2 *Algorithm Performance*

To evaluate the performance of different prediction algorithms, we conducted a comparison experiment. Table 2 presents the performance results with different features applied to the prediction algorithms. Specifically, in addition to the three types of features we introduced in Section 2.2, we also consider the combination of market attractiveness and market competitiveness features (at $K = 10$) as *review-based* features, and the combination of all the three different types of features (i.e., market attractiveness features, market competitiveness features and geographic features). From Table 2, we can see that SVR algorithm achieves slight improvement against Ridge regression. For example, in terms of Spearman's ρ, SVR algorithm obtains 0.5% improvement with geographic features and 1.5% improvement with review-based features. In addition, we can observe that GBRT algorithm performs better than Ridge regression and SVR algorithms with respect to all features. For example, when only considering the review-based features, GBRT algorithm achieves 6.4% improvement relative to SVR algorithm ($RMSE = 0.597$ versus $RMSE = 0.638$).

	RMSE	Spearman's ρ	MAP
Geographic Features (GeoF)			
Ridge	0.795	0.187	0.0011
SVR	0.793	0.188	0.0014
GBRT	**0.761**	**0.192**	**0.0017**
Market Attractiveness Features (MAF)			
Ridge	0.657	0.215	0.0023
SVR	0.642	0.223	0.0021
GBRT	**0.604**	**0.241**	**0.0025**
Market Competitiveness Features (MCF)			
Ridge	0.772	0.200	0.0015
SVR	0.768	0.203	0.0015
GBRT	**0.760**	**0.205**	**0.0016**
Review-based Features (MAF & MCF)			
Ridge	0.646	0.221	0.0024
SVR	0.638	0.235	0.0021
GBRT	**0.597**	**0.244**	**0.0025**
Combined Features (GeoF & MAF & MCF)			
Ridge	0.631	0.236	0.0026
SVR	0.620	0.251	0.0028
GBRT	**0.586**	**0.308**	**0.0031**

Table 2: Overall performance comparison of different prediction algorithms in terms of $RMSE$, Spearman's ρ and MAP.

3.3.3 *Feature Performance*

From Table 2, we can also observe that the predictions based on features mined from reviews are more accurate than those based on geographic features. For example, by using the GBRT algorithm, the market attractiveness features achieve 20.6%, 25.5% and 47.0% improvements relative to geographic features in terms of RMSE, Spearman's ρ and MAP respectively. What's more, the market attractiveness features obtain better results than the market competitiveness features across all measures. It can also be seen that the combination of market attractiveness features and market competitiveness features (i.e., review-based features) performs better than geographic features in terms of all measures. Moreover, the combination of all features achieves better results than taking geographic and review-based features alone with regard to $RMSE$, Spearman's ρ and MAP@10.

4. RELATED WORK

The optimal placement problem, which is also called the site selection problem [16], has attracted research attention in recent years. A common approach to assessing location is to first develop a checklist to ensure that all relevant factors are considered [2]. However, some of the factors may be quite subjective, and the acquisition of them typically depends on certain local demographics, such as transportation links, and the ease of ingress and egress in the area, which are expensive and time-consuming to acquire.

There are several spatial interaction approaches that draw on the assumptions that the intensity of between-location interactions decreases with distance and the quality of a retail store location increases with the intensity of use and proximity of complementary locations [3, 12]. However, these approaches are applicable only to the selection of agglomeration locations, such as for large shopping centres.

Some researchers have proposed approaches based on analysis of the spatial distribution of commercial activities [8, 11, 19]. For example, in [11], the human mobility trace (check-ins) data in location-based social networks (LBSN) are used to mine certain features in order to provide insights into the quality of an area for opening a new restaurant. However, the limitation of this work is that only three chain restaurants are evaluated (Starbucks, Dunkin' Donuts and McDonald's). Its applicability to other restaurant chains is unclear. In our work, we particularly investigate whether the information embedded in user-generated reviews can be helpful for solving the optimal placement problem. To our knowledge, this is one of the first efforts to tackle the problem by leveraging the publicly available user-generated reviews.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we study the problem of optimal restaurant placement, by taking advantage of online user-generated reviews. We design three sets of features, namely, review-based *market attractiveness* features, review-based *market competitiveness* features, and *geographic* features, which are then exploited to predict the potential number of visits for a prospective restaurant in a given location. We also conducted an experiment with real-world restaurant data to investigate the predictive power of our constructed features. In the experiment, we evaluated each set of features separately, and found that the market attractiveness and market competitiveness features have greater predictive value than geographic features. The prediction accuracy is further increased when both geographic and review-based features are considered. We can hence conclude that the information embedded in user-generated reviews can be helpful to tackle the optimal restaurant placement problem.

In the future, we plan to: (i) exploit additional types of features to improve prediction accuracy (e.g., traffic conditions); (ii) evaluate the effectiveness of our method with other datasets; (iii) show some real case study in our experiments; and (iv) develop algorithm to identify the optimal location for a new restaurant.

Acknowledgement

We thank Hong Kong RGC and China NSFC for sponsoring the described research work (under projects RGC/HKBU12200415 and NSFC/61272365).

6. REFERENCES

[1] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proc. 25th SIGMOD*, pages 49–60. ACM, 1999.

[2] W. Applebaum. Can store location research be a science? *Economic Geography*, pages 234–237, 1965.

[3] A. Athiyaman. Location decision making: The case of retail service development in a closed population. *Academy of Marketing Studies*, 15(1):13, 2010.

[4] O. Berman and D. Krass. The generalized maximal covering location problem. *Computers & Operations Research*, 29(6):563–581, 2002.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[6] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proc. 5th LREC*, pages 417–422. 2006.

[7] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[8] A. S. Furtado, R. Fileto, and C. Renso. Assessing the attractiveness of places with movement data. *Journal of Information and Data Management*, 4(2):124, 2013.

[9] N. Hing. Franchisee satisfaction: Contributors and consequences. *Journal of Small Business Management*, 33(2):12, 1995.

[10] C. H. Hsu, T. F. Powers, and T. F. Powers. *Marketing Hospitality.* John Wiley & Sons, 2002.

[11] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proc. 19th KDD*, pages 793–801. ACM, 2013.

[12] A. Kubis and M. Hartmann. Analysis of location of large-area shopping centres. a probabilistic gravity model for the halle–leipzig area. *Jahrbuch für Regionalwissenschaft*, 27(1):43–57, 2007.

[13] M. Mehaffy, S. Porta, Y. Rofè, and N. Salingaros. Urban nuclei and the geometry of streets: The 'emergent neighborhoods' model. *Urban Design International*, 15(1):22–46, 2010.

[14] S. Moghaddam and M. Ester. On the design of LDA models for aspect-based opinion mining. In *Proc. 21st CIKM*, pages 803–812. ACM, 2012.

[15] J. L. Myers, A. Well, and R. F. Lorch. *Research Design and Statistical Analysis.* Routledge, 2010.

[16] T. R. Rex and K. S. Walls. Site selection factors vary widely by economic cluster. *Arizona Business*, pages 6–8, 2000.

[17] A. Smola and V. Vapnik. Support vector regression machines. *Advances in NIPS*, 9:155–161, 1997.

[18] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *Proc. 17th WWW*, pages 111–120. ACM, 2008.

[19] M. Xu, T. Wang, Z. Wu, J. Zhou, J. Li, and H. Wu. Store location selection via mining search query logs of baidu maps. *arXiv preprint arXiv:1606.03662*, 2016.