A Multi-Objective Approach for Optimal Store Placement

Jeroen Rook
LIACS, Leiden University
j.g.rook@umail.leidenuniv.nl

Brent Verpaalen
LIACS, Leiden University
b.a.a.verpaalen@umail.leidenuniv.nl

Daniela Gawehns
LIACS, Leiden University
d.gawehns@liacs.leidenuniv.nl

Mitra Baratchi
LIACS, Leiden University
m.baratchi@liacs.leidenuniv.nl

***Abstract* - When a store owner wants to open a new store he or she desires a location that attracts a large number of customers. Previous work has shown how location based social networks can contribute to this decision process. However, opening a store also has an impact on the surrounding neighbourhood. With the use of urban planning theories we define a score showing the impact of these store placements. We propose a framework, that selects the best venue categories for a given location in a city according to their scores for both perspectives. These scores are computed from metrics extracted using a location based social network from Foursquare. Our experiments, based on the city of New York, show that the number of suitable store categories, for a single location, are often not singular. This indicates that this multi-objective approach is necessary in solving the optimal store placement problem.**

1. INTRODUCTION

Imagine you want to open a new restaurant. Where in the city would you allocate your restaurant? As a store owner you desire a location where the characteristics of the surrounding area give you a large number of customers, resulting in an increased revenue. However, store owners need permissions from local governments, who also consider urban planning objectives of the city. These objectives are set out to maintain a liveable and economically thriving city. As a store owner it is beneficial to take these objectives into account. These objectives provide a sustainable future for the store, preventing bad cityplanning, which could be disastrous for the cities economy [1]. Earlier research used Foursquare data sets in order to predict the expected number of visitors [2]. With the addition of the governmental view on store placement we extended the optimal store placement problem. Approaching the placement problem, to our knowledge, has not been covered in previous research.

In this paper, we rewrite the optimal store placement problem by trying to find the location that is best for both the store owner and the city itself. This is done by mining features from a location based social networks data set, provided by Foursquare. We use known features, and also introduce two new features derived from a relative neighbourhood graph. With these features we predict the popularity of a location and define an objective function representing the impact on the city. These two functions allow us to obtain a Pareto optimal set of the most suitable categories for a given locations, helping potential store owners to obtain popular store locations and local governments in understanding which venue categories should be targeted for a specific location.

In this paper we focus on the following contributions:

- Creating a multi-objective approach on location based store category prediction;
- Constructing new features from a relative neighbourhood graph in order to capture the function of the neighbourhood within the city;
- Applying our proposed method to the problem of identifying optimal store locations in New York City using a Foursquare movements network.

2. RELATED WORK

Traditionally, the optimal store placement problem revolved around the central place theory, the spatial interaction theory and the theory of minimal differentiation [3]. With the rise of location based social networks (LBSNs), which combines data from social networks with real-world objects, other approaches in solving the placement problem arised. Karamshuk et. al. [2] defined this problem as a ranking problem for a set of locations using one particular type of store. Another proposed approach is to make a prediction where a category would be ranked by its popularity at a given location using matrix factorisation [4]. Other socially generated data such as written visitor reviews have shown to be able to contribute in solving this problem [5, 3]. The focus of these works are based on the optimality from a store owners perspective, which is defined as a maximum number of visitors.

Considering this problem from the point of view of urban planners is much wider than only the number of visitors, but cannot be defined on one single property. Proposals of definitions have been made [6] and commonly fall back on traditional theories. One of these theories is mixed-land use [7], which sets out to have residential, commercial, and working locations within the same neighbourhoods. This encourages non-auto commuting, which accounts for a decrease of traffic congestion, and results into less CO2 emissions. Both having a positive effect on the city.

3. METHODS

Before the most suitable venue category for a given location can be found, several steps need to be taken. Our proposed approach is performed in two phases, training and ranking. In the training phase, features for each location in the data set are constructed, and the number of check-ins per venue are calculated. These check-ins are used as our ground truth for a regression model predicting the number of check-ins from the given features. In the ranking phase, for any given location we artificially simulate a venue category placement, computing resulting features. This is done for all possible categories. For these features the predicted popularity is retrieved using the regression model created in the training phase. Also the city impact score for each of these categories is computed. In the

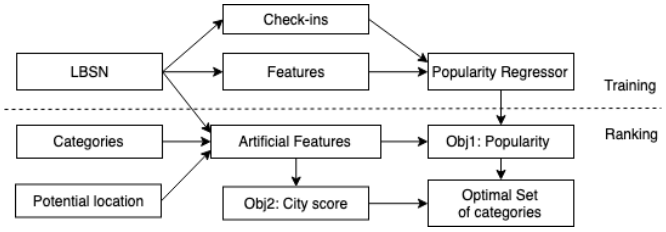

Figure 1: The full framework, for getting multiple objective rankings comparing different categories at a given location. These rankings are used to compute an optimal set of categories.

end we can obtain the Pareto optimal set of categories for that particular location, based on the two objectives. Categories in the Pareto optimal set have the best pairs of scores. All these steps and their underlying dependencies are visualised in Figure 1.

3.1 Feature engineering

In order to make rankings on popularity and city impact we hand-craft features, for each location per category, from the LSBN. We make heavy use of the features used by Karamshuk et al. [2], which are based on the work of Jensen [8]. Karamshuk et. al. extracted a total of eight different features from the location based social network, which are the first eight described features in Table 2. These features are divided in two different groups. The first group are features constructed from the geographical situation and the other group are features constructed using the movements between venues, which are called mobility. We extend this feature set by looking at the geographical behaviour of the movements in the LSBN. These features are described in 3.1.1.

All of the Karamshuk et al. features are based on the interactions occurring in the direct neighbourhood of the venue under investigation. A neighbourhood is defined as the area within a circular radius r, which is set to $200m$ for all features. The Haversine distance [9] is used as a distance metric. Due to lack of space, we refer to their paper for the full details on the construction of the features.

We will now outline the two new features Bypass and Betweenness centrality:

3.1.1 FlowGraph

The existing features do not capture the the function of the neighbourhood within the city as a whole. For example, it does not show what paths people potentially took to get from one venue to another. To get a better understanding of this we create a relative neighbourhood graph (RNG) [10] from all venues. This graph only has edges between neighbours from a geographical perspective. An edge between the venues (a, b) only appears if the following holds:

$$\forall c \in V : dist(a, c) > dist(a, b) \wedge dist(c, b) > dist(a, b) \quad (1)$$

Where V is the set of venues in the RNG and $dist$ is the Haversine distance between two venues. When all eligible edges, according to equation 1, are defined in the graph, the movements from the LSBN are mapped to corresponding edges of the shortest path between the venues in the RNG. The obtained weighted graph is referred to as FlowGraph, which shows dense areas and areas which acting as hubs for the whole graph. As a byproduct it allows for a visualisation of the interaction structure in the city, as can be seen in Figure 3. Features constructed from this graph are grouped under *Geographical Flow*.

Table 1: Overview of the constructed features extracted from the venue movements network

Feature	Description
Geographical	
Density	Number of neighbours
Neighbours Entropy	Spatial heterogeneity
Competitiveness	Amount of same category venues in the neighbourhood
Quality by Jensen	Amount of categories which tend to occur together
Mobility	
Area Popularity	The number of check-ins in the neighbourhood
Transition Density	The number of check-ins of venues within the neighbourhood
Flow	The number of check-in from movements outside the neighbourhood
Transition Quality	The expected number of movements from venues in the neighbourhood based on their categories.
Geographical Flow	
Bypass	Indegree of bypassing users
Betweenness centrality	Amount of times the location is on a shortest path

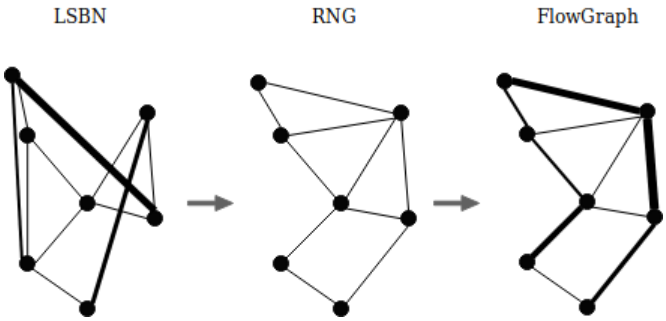

Figure 2: Sequential construction of the FlowGraph, which is build by creating a relative neighbourhood graph based on the venue locations and mapping the weights of the edges from the LSBN to all edges on the shortest paths in the RNG.

The *bypass* feature is the indegree of a location in the FlowGraph. A high value suggests that many people are passing through the selected location. This is especially useful in less occupied neighbourhoods, since people passing by are potential customers.

Another feature we propose is *betweenness centrality*, which was not considered in [2]. This centrality is defined by the frequency a node appears in the shortest paths between all nodes in the graph. This is related to the hub function for each location. For example the area which connects Long Island with Manhattan has the highest betweenness centrality.

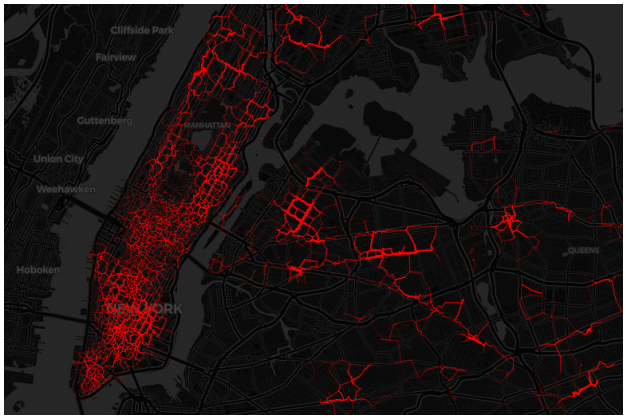

Figure 3: The FlowGraph for the venues of New York City. Busy areas and corridors can quickly be identified.

3.2 Objectives

In order to capture the venue categories that are most suitable considering both the store owners and the city perspectives we define an objective for each of them. Thus, for each category at a given location, two scores are computed. A category is covered if and only if there exists another category for which all scores are higher than it's own. From the set of categories we define the optimal set as the ones that are not covered.

3.2.1 Popularity

The benefit for the store owner is defined in terms of popularity, i.e. check-ins the venue would potentially receive. For each location we only know the popularity for the category, which is already there. For all other categories at a location this is unknown. A regression model is trained on the existing venues with their category in the data set. The model predicts for a given category and corresponding extracted features, as described in the previous section, what the popularity probably will be. With this model we can get a better understanding of how popular the venue would become for all other categories at a location. The popularity is the objective for the store owners and should be maximised.

3.2.2 City impact score

We investigated what objectives could be important from the city perspective. We found one important feature that can be extracted from data, the *Neighbours Entropy* [11, 2]. This feature represents the diversity of categories within a neighbourhood, which is in essence mixed-land use, the criteria which is regularly considered by urban planners. When different categories are given for a location the entropy changes accordingly. A high entropy means a high diversity of the surrounding, so this score should be maximised.

4. EXPERIMENTS

In this section we perform a study on the Foursquare data acquired from the city of New York. We apply our proposed methods for the city of New York in order to see if the extracted features from the FlowGraph significantly contribute towards better popularity predictions. Furthermore, we show the results for the optimal set of categories. Before these experiments can be conducted the data set is pre-processed to a suitable format.

4.1 Data set

The data set consists of two different data structures. The first structure holds information of all the known venues in the city, such as venue name, longitude, latitude, and the venue category. The other structure consists of aggregated movements between two venues. The provided movements are between two venues within the city, and also includes movements going to or leaving a single venue in the city. We defined the number of check-ins. i.e. popularity, as the total number of movements going towards a venue.

In total there are $17,382$ venues for the city of New York where at least one check-in occurred. The number of distinct categories is 503, which is quite specific. To be sure this would not result in an insufficient number of training examples per category, we clustered multiple categories to prevent overfitting on a handful of venues. To do this we reduced the categories to their top-level category from Foursquare, resulting in a total of 10 categories: *Arts & Entertainment, Shops & Services, Professional & Other Places, Food, Residences, Travel & Transport, College & Universities, Outdoors & Recreation, Nightlife Spots, and Events.*

4.2 Popularity Prediction

We created venue popularity predictions for locations using different venue types and their corresponding features. For this predictions we choose to use a regressor because of the continuous property of our popularity variable. A Catboost regression model [12] is used for getting popularity predictions. Catboost uses gradient boosting on decision trees. We evaluated the performance of the algorithm using the root mean square error metric. The data set was split into a train and test set with a distribution of respectively 80 and 20 percent. The model was trained on the train set and the test set was used for evaluation, preventing data leakage. When training our algorithm we got the first row of results shown in Table 2

Table 2: Error scores (RMSE) of the popularity regressions models provided with different feature sets

	Train	Test
All features	129.18	128.54
Original features	129.28	128.65

In order to see if our constructed features improved the popularity prediction we performed the test twice using different feature sets. The first set included all described features, the second set contained only the original features excluding ours. The results show that the added features in our situation have low to no impact on the popularity prediction error. A paired t-test between predictions from both sets yielded a p-value of 0.7. This means that our proposed features do not create a significant difference in performance.

4.3 Optimal sets

From all venues in the data set we took a random sample of 1000 venues. We created features for each category available at all locations. Each location we analysed the resulting category sets. Recall that each category is assigned with two scores. As an example the categories are projected on a 2D plot where each score is on one of the axis, as can be seen in Figure 4. In the original dataset this is a Taco Shop in the Upper East Side. In total there are four categories in the optimal set, highlighted in red. These categories are thus to

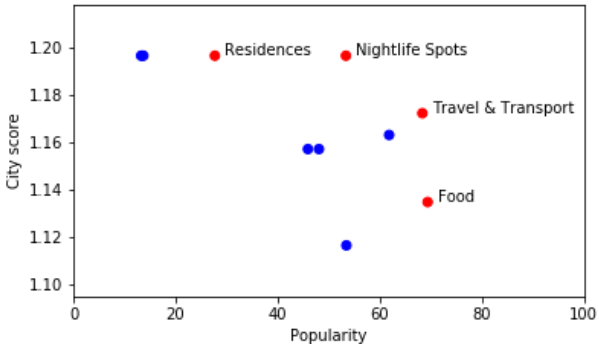

Figure 4: Projection of all categories for a location, currently used as a Taco store in the Upper East Side, with the optimal venues shown in red.

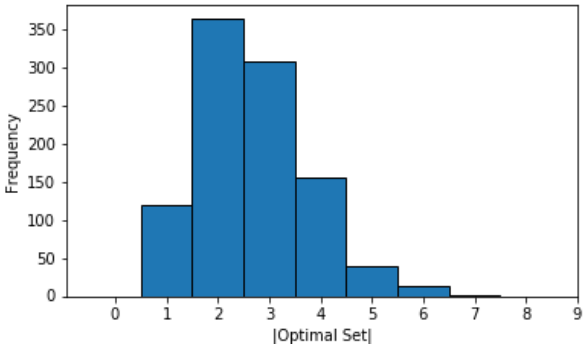

Figure 5: Histogram of the number of categories in the optimal set from a sample of 1000 locations

be considered as most suitable for that particular location. In this set *residential* and *nightlife spots* would give a higher diversity in the area and a food venue is suggested be most popular.

The histogram at Figure 5 shows the distribution of categories in the optimal set for all locations in the sample. A majority of locations have more than one category in their optimal set. Meaning that there often is a conflict between the selection of the best category between each score. This conflict shows us that the venue prediction problem is a multi-objective problem. Because a large majority has multiple optimal venue categories we know that it is hard to find the best category when only looking at one objective.

5. CONCLUSIONS

For future work we want to take an in depth look into other city planning objectives. Currently, we are focused on the shop diversity as a city planning objective. City planning is more than just this feature, but is also often a city specific task. If we want to research this, we should create a closer co-operation with the cities we are analysing, increasing the effectiveness of our city score. Another addition is taking more different venue categories into account. We focused on a generalised set of categories. If we could gather more data from different venues we could look at a lower type in the Foursquare venue type hierarchy. One of the proposed goals could be to analyse different food type venues, such as but not limited to: fastfood, fine dining, coffee shop, etc.

In conclusion, using a Foursquare movements network we created a framework where the optimal store placement problem can be scored using the perspective of the store owner, and the perspective of the city as a system. We introduced a projection of movements on a relative neighbourhood graph, which we called FlowGraph, and used this graph to create two new features. The impact of these features were minimal and not significant for the performance of popularity prediction. Overall we showed that both objectives rank the categories in a different way, resulting in multiple categories considered optimal. This indicates that the multi-objective approach is a necessity in order to work towards a city improvement based on city planning guidelines while having satisfactory store owners.

6. REFERENCES

[1] M. Townsend, J. Surane, E. Orr, and C. Cannon. America's" retail apocalypse" is really just beginning. *URL: https://www.bloomberg .com/graphics/2017-retail-debt/*, 2017.

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

[3] H. Damavandi, N. Abdolvand, and F. Karimipour. The computational techniques for optimal store placement: A review. In *International Conference on Computational Science and Its Applications*, pages 447–460. Springer, 2018.

[4] Z. Yu, M. Tian, Z. Wang, B. Guo, and T. Mei. Shop-type recommendation leveraging the data from social media and location-based services. *ACM Trans. Knowl. Discov. Data*, 11(1):1:1–1:21, July 2016.

[5] Y. Fu, G. Liu, S. Papadimitriou, H. Xiong, Y. Ge, H. Zhu, and C. Zhu. Real estate ranking via mixed land-use latent models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 299–308. ACM, 2015.

[6] A. Gil Solá and B. Vilhelmson. Negotiating proximity in sustainable urban planning: A swedish case. *Sustainability*, 11(1):31, 2019.

[7] R. Cervero. Mixed land-uses and commuting: Evidence from the american housing survey. *Transportation Research Part A: Policy and Practice*, 30(5):361–377, 1996.

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

[9] C.C. Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

[10] J.W. Jaromczyk and G.T. Toussaint. Relative neighborhood graphs and their relatives. *Proceedings of the IEEE*, 80(9):1502–1517, 1992.

[11] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[12] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, pages 6638–6648, 2018.