Where is Beverly Hills in your town? Finding similar neighborhoods across cities through social media activity

Géraud Le Falher[1]
Inria Lille, France
geraud.lefalher@inria.fr

Aristides Gionis
Aalto University, Finland
aristides.gionis@aalto.fi

Michael Mathioudakis
Helsinki Institute for Information Technology (HIIT)
michael.mathioudakis@hiit.fi

Given a neighborhood in one city, we want to find in others cities the most similar neighborhood in terms of human activities. Such answers improve touristic recommendation and urban planning. We harness Foursquare and Flickr data collected in 20 cities to learn ad-hoc distances, combine them into an efficient search algorithm and show its compelling results.

1 Introduction

Say you are living in a lovely, green and quiet neighborhood of Paris but are looking for a hotel in Helsinki. Or imagine that, working at a municipality, you are faced with discontent in a localized area, but cannot find any explanation for it. In both scenarios, you will benefit from a method that takes as input a neighborhood in a city you know and presents you with similar locations elsewhere in the world, whether to search for accommodation or to discover comparable issues and their potential solutions. Whereas already possible by asking carefully selected relatives, the wealth and breadth of data shared by people living in smart cities makes it more affordable, tough requiring meticulous implementation.

Our method considers neighborhoods as sets of venues so we first gathered data about venues and how people interact with them from now ubiquitous social media (2.1). Given a query neighborhood, we look in other cities for sets of adjacent venues sharing similar features and thus hosting similar activities. We solve this problem in 3 stages:

- First we learn distances between venues and evaluate them in an information retrieval fashion (2.2)
- Then we collected neighborhoods ground truth and evaluate how different distances between set of venues are able to recover it (2.3)
- Finally we design a heuristic to quickly search through all possible neighborhoods (3.1) whereas still providing consistent results (3.2)

We conclude by reviewing other solutions and discussing possible extensions of our methods (4).

2 Data and methods

2.1 Dataset Our notion of similarity aims to match human perception, and that guides us in the choice of data source. For instance, Foursquare is a location based social network that lets its users announce to the world when they *check-in* in various *venues*. In addition to traditional hotels and airports, venues include places of all kind like restaurants, parks or museum. By crawling check-ins from Twitter between March and July 2014 and using a previous dataset [2], we end up with 5 million individual data points. We also pulled information from Foursquare regarding 87,000 venues referenced by more than five check-ins. This includes their location, their category and the total number of visits. Lastly, we collected 8 million of timestamped, geolocalized photos shared on Flickr to better quantify activity level through space and time.

[1]Work done while working at Aalto University.

Figure 1: 2D embedding of 23086 venues from 10 European cities.

We aggregate these data at venue level to describe venues by their popularity, the time of day/week at which they are active, the diversity of their audience, the density of their surrounding and so on. To illustrate the expressiveness of these features, consider Figure 1, a dimensionality reduction computed by t-distributed Stochastic Neighbor Embedding (t-SNE) [7]. Hiding the true category, venues of the same kind are still projected together: Education on the right, Nightlife on the bottom left, Professional on the bottom right or Recreation on the top.

2.2 Distance between venues Once we represent venues by such numeric feature vectors, we can compute distance between them, for instance with standard Euclidean distance. Yet we would like to learn distances that bring similar venues closer to each other while pulling apart dissimilar ones [1]. We experiment with two well-established methods: (*i*) Information Theoretic Metric Learning (ITML) [5] and (*ii*) Gradient Boosted Largest Margin Nearest Neighborhood (LMNN) [6]. Furthermore, we project venues on a 2d plane using (*iii*) t-SNE and compute distances in this reduced space.

Because we do not have explicit information about venues similarity, we resort to indirect labels during the semi-supervised learning phase. Namely, we deem venues to be similar if they belong to the same Foursquare top-level category and dissimilar otherwise.

We devise two tasks to evaluate how well these distances express similarity between venues. First, we pick a venue of a global brand (like *Starbucks* or *McDonald's*) in one city and ranked all venues of another city using the distance under consideration. Then we look how far are venues of the same brand in this ranking. Second, we pick a venue of a given subcategory (say *Italian restaurant*) and see how close are venues of the same subcategory in another city. We find

that, while LMNN performs better, the difference with the EUCLIDEAN distance is not as marked as expected; probably because our labeling is noisy.

2.3 Distance between neighborhoods Next we need to evaluate distance between two sets of venues forming neighborhoods. We use the Earth Mover's Distance (EMD), which measures the *total amount of work* needed to transform (move) one vector set (total mass) to the other [9]. It is parametrized by the underlying distance between individual vectors, for which we use the four metrics mentioned earlier.

Because we cannot evaluate the distance of the exponential number of possible neighborhoods, we restrict ourselves to circles of various radii, centered over a regular grid paving the target city. We returned the ones that are the closest to the query neighborhood.

To assess the relevance of our results, we first pick 8 thematic neighborhoods in Paris, for instance the *16th arrondissement*, which is home of upper-class families and where real estate is expensive. Then we ask acquaintances to give us comparable areas in the city they live in and know well. It turns out that as EMD underlying distance, EUCLIDEAN is again the best, returning neighborhoods that overlap the most with the ground truth provided by our local experts.

3 Faster search and its results

3.1 Heuristic search The exhaustive search described earlier is computationally expensive, taking more than 30 minutes for a complete scan of New-York on 4 cores. Thus, we develop an alternative strategy. It starts by computing pairwise distances between (*i*) the venues forming the query neighborhood and (*ii*) all the venues in the target city. Then we prune the search space in the target city by keeping only a fraction of its venues that are the closest to those in the query neighborhood. Assuming the answer will be an area dense of these "anchor" venues, we cluster them using DBSCAN. To account for missed venues, we obtain candidate neighborhoods from these clusters by expanding their border three times. By restricting ourselves to these promising areas, we perform much fewer expensive EMD evaluations.

3.2 Experimental results In our experiments, we measure that this heuristic search is between 10 and 1000 times faster than the exhaustive one. Yet when comparing the distances of the best result returned by both methods, we find that heuristic search is better half of the time, as it finds arbitrarily shaped neighborhoods. To illustrate this, we present a qualitative example in Figure 2: pricey real estate in Washington and New-York. Bethesda, one the most affluent suburb in the US, was identified by our expert as fitting this description (shown in orange) and we use it as our query. The most similar neighborhood suggested by our method (shown in blue) overlaps significantly with the surrounding of the Fifth Avenue, indeed one of the most expensive street in the world and the designated ground truth (in orange).

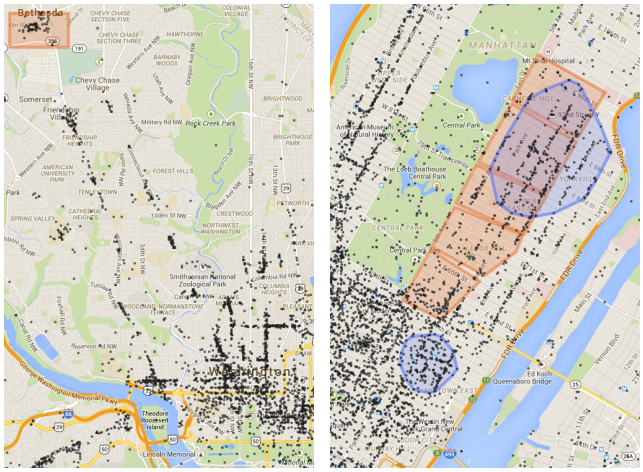

Figure 2: A query in Washington and its result in New-York.

4 Related work and discussion

Fueled by the explosion of social media data, urban computing is a lively field, and two recent works exploit Foursquare check-ins to better understand cities, although they are more concerned about segmenting cities into regions called respectively Livehoods [4] and Hoodsquare [10] rather than computing similarities across cities. Another approach to this clustering task is to perform topic modelling [3].

Our flexible approach would benefit from incorporating this finer categorisation, as well as sentiment analysis from textual data such as tweets and venue reviews. Another extension would be to automatically identify neighborhoods from the data and match them to compute similarities between cities as a whole [8].

References

[1] A. Bellet *et al.*, "A survey on metric learning for feature vectors and structured data", Université de Saint-Etienne, Tech. Rep., 2013.

[2] Z. Cheng *et al.*, "Exploring Millions of Footprints in Location Sharing Services", in *ICWSM*, 2011.

[3] J. Cranshaw *et al.*, "Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with Latent Topic Modeling", in *NIPS CSS*, 2010.

[4] J. Cranshaw *et al.*, "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City.", in *ICWSM*, 2012.

[5] J. V. Davis *et al.*, "Information-theoretic Metric Learning", in *ICML*, 2007.

[6] D. Kedem *et al.*, "Non-linear Metric Learning", in *NIPS*, 2012.

[7] L. V. D. Maaten *et al.*, "Visualizing Data using t-SNE.", *JMLR*, 2008.

[8] D. Preoţiuc-Pietro *et al.*, "Exploring venue-based city-to-city similarity measures", *UrbComp*, 2013.

[9] Y. Rubner *et al.*, "A metric for distributions with applications to image databases", in *ICCV*, 1998.

[10] A. X. Zhang *et al.*, "Hoodsquare: Modeling and Recommending Neighborhoods in Location-Based Social Networks", in *SocialCom*, 2013.