Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with Latent Topic Modeling

Justing Cranshaw
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
jcransh@cs.cmu.edu

Tae Yano
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
taey@cs.cmu.edu

Abstract

We report our initial experiments on a computational approach to neighborhood detection from community-authored location annotated content. Automated discerning of context from the physical environment is an important challenge in the age of mobile computing, where *context awareness* can often be a critical component in the design of pervasive applications. Furthermore, defining the context of a location, or a "neighborhood," in an urban landscape is often a difficult task, since neighborhoods are often defined by intangible qualities such as shared function or community structure, rather than the prescribed geopolitical jurisdictions. In this study we seek to develop a data-driven definition of a neighborhood using topic modeling and socially annotated location data. We treat the question as a matter of latent topic discovery. Using incidental, geo-coded data gathered from location-sharing social networks, we show that we can discover canonical neighborhoods by analyzing co-occurence patterns in place categories. We report a qualitative evaluation of our approach, and interpret the validity of the discovered prototypical neighborhoods. We also discus future work and potential implications of our approach to the social sciences.

1 Introduction

The notion of neighborhood is intrinsic to urban dwellers, as neighborhoods can be an important factor that drives and constrains the everyday lives of their inhabitants. However, the precise definition of neighborhood, is rather ineffable. Two neighbors who agree on the unique assets of their neighborhood can often disagree on exactly where their neighborhood begins and ends, since there are often no reliable defining boundaries. Functional neighborhood demarcations often do not quite align with the geo-political jurisdictions, such as the zip code in the United States. Furthermore topological features such as rivers and valleys are not always a decisive factors in an urban landscape. There may not be one right answer to this problem.

Despite this ambiguity, the analysis of neighborhoods has always been a keen interest of businesses such as realtors, developers, and urban planners. To them the "quality" of a neighborhood significantly effects the value of their assets. The optimal pricing of a house, the best strategic location of a new franchise, or the optimal routing of a new bus line are all unknowable without first analyzing the neighborhoods of interest. In an academic setting, the analysis of neighborhoods involves much more than a monetary appraisal. Researchers in social phycology, political science and public health often study "the neighborhood effect," which aims to understand neighborhood and community level factors that influence observable phenomena such as obesity rates, perceived happiness and voting patterns.

In ubiquitous computing research, physical location has attracted a new surge of interest in recent years due in-part to the advent of the context-aware computing paradigm, where researchers and practitioners seek to develop systems that not only adapt to the subjects' *personal* actions and dispositions but also to the *physical* context of their environment [4]. A related interest in location has come from the rise of location-based services, in particular location-sharing social applications such as foursquare (http://foursquare.com), Yelp (http://www.yelp.com) and Facebook Places (http://www.facebook.com), as well as location-sharing research projects such as Locaccino (http://www.locaccino.org). This rise of location-sharing platforms has created promising new opportunities for computational social science; Researchers now have a channel to collect geo-coded traces of user activities, making it possible to ground human behavior to the physical environment on a large scale. These new technologies have also created the need to better understand the relationship between privacy and location. Automated tools that aid in understanding neighborhood functionality could be used to help users control the information that they wish to share with such systems [7].

In this paper, we report our initial experiments in using latent topic modeling to analyze a collection of geo-coded augmented location category tags which were generated as a part of social tagging in a location sharing application. Our goal in this work is to distill a set of prototypical descriptions of the "sense of neighborhood" from this incidental data in a simple and intuitive manner, while avoiding human supervision as much as possible.

2 Topic modeling for proto-neighborhood discovery

Throughout the course of the study, we take an utilitarian approach to the definition of neighborhood; We hypothesize a "neighborhood" to be a distinctive, discernible, co-occurrence pattern in functional descriptions of the places within a region. In this sense what we are after is the canonical forms, or archetypes, of neighborhoods. This is a convenient assumption from the perspective of context-sensing, since in this canonical, lower dimensional landscape, there are much fewer neighborhood contexts to keep track of, which aids in the understandability of the end result.

Topic modeling is a widely used technique, most often seen in text analysis and in clustering of text collections. Here we apply the technique to tease out the co-occurrence patterns in user generated venue descriptions. Our hypothesis is that, if coherent patterns, prototypes, are repeated enough throughout a landscape, the topic model will be able to tease them out as a distinctive set of distributions over the venue types.

2.1 Data Collection and Processing

Investigating these hypotheses requires descriptive data about the types of places in many distinct geographic regions. To meet this need, we gathered data from foursquare, a popular location-based social network that allows users share their location with their friends by "checking-in" to the places that they visit. When users check-in to a newly added venue[1] they can provide descriptive data about the venue to foursquare, including a venue category, which is chosen from a fixed set of categories.

We note that there are other sources for such venue categorization data (Google and Yelp for example). We chose to use foursquare in our experiments for two reasons. First, categorization in foursquare is hierarchical. Every venue is described by one of 8 top level categories: *Arts & Entertainment*, *College&Education*, *Food*, *Home/Work/Other*, *Nightlife*, *Parks&Outdoors*, *Shops*, and *Travel*. The description of each venue is then further refined by a sequence of subcategories, so that each venue is categorized by a path in the category hierarchy. For example a wine shop is categorized as *Shops : Food&Drink : Wine Shop*. Though we do not make explicit use of the hierarchy in this work, we anticipate future models might benefit from the use of this added structure to the data. In addition to the hierarchical categorization, we found that the root categorizes of venues were well organized, concise and descriptive, which was particularly appealing for our task. In total, there are currently 347 root foursquare venue categories.

Using the foursquare API, we downloaded descriptions of 494,732 venues distributed over 12 metropolitan areas in the Unites States, including New York City, Los Angeles, San Francisco,

[1] The places where users can check-in to in foursquare are called *venues*. We inherit this terminology.

Topic	Top categories
0	Home, Pizza, Salon/Barbershop, Beauty/Cosmetic, Bank/Financial, Other - Buildings, Food:Chinese
1	Doctor's Office, School, Medical, Dentist's Office, Hospital, Religious, Gym/Fitness
2	Boat/Ferry, Golf Course, Harbor/Marina, Farm, Campground, Lake/Pond, Arcade
3	Pub, Resort, Community College, Racetrack, Motel, Beer Garden, Hotel Bar
4	Beach, Seafood, Scenic Lookout, Surf Spot, Harbor/Marina, Landmark, Skate/Surf/Snowboard
5	Parks&Outdoor, Entertainment, Baseball Field, Theme Park, Zoo/Aquarium, Baseball, Stadium
6	Hiking Trail, History Museum, Hostel, Post Office, Casino, Parks&Outdoor, Scenic Lookout
7	Bar, Event Space, Art Gallery, Entertainment, Discotheque, Taxi, Speakeasy/Secret Spot
8	Corporate/Office, Home, Coworking Space, Buildings, Tech Startup, Gym, Hotel
9	Grocery/Supermarket, Flower Shop, Drug Store, Post Office, Italian, Flea Market, Breakfast/Brunch
10	Boat/Ferry, Golf Course, Harbor/Marina, Farm, Campground, Lake/Pond, Arcade
11	Automotive Shop, Mexican, Asian, Chinese, Vietnamese, Korean, Hardware
12	Airport Gate, Plane/In-flight, Plane, Airport, Terminal, Other - Travel, Hotel
13	Fast Food, Mexican, Gas Station, Automotive Shop, Pizza, Other - Buildings, Burgers
14	Other - Buildings, Government, Train Station, Bus, Light Rail, Courthouse
15	Highway/Traffic, Golf Course, Bridge, Farm, Cemetery, BBQ, Field, Subway
16	Other - Travel, Coffee Shop, Truck/Street Food, Sandwiches, Garden, Steakhouse, Karaoke
17	Home, Playground, Pool, Speakeasy/Secret Spot, College&Education, Gym
18	Park, Dog Run, Playground, Parks&Outdoor, Pool, Music Venue, Skate Park, Classroom
19	Church, High School, Field, College&Education, Academic Building, Classroom, Library
20	Gas Station/Garage, Pool, Playground, Diner, Automotive, Financial, Fire Station
21	Winery, Bed&Breakfast, Wine Bar, Vineyard, New American, Wine Shop, Resort
22	Apparel, Department Store, Shopping , Women's Apparel, Shoes, Cosmetic, Furniture
23	Academic Building, Dorm, Administrative Building, Classroom, Library, University
24	American Food, Food:Other, Pizza, Mexican, Italian, Burgers, Steakhouse

Table 1: The most probable words from the learned topic distributions over venue categories.

Seattle, Boston, and San Diego. For each venue the API provides a venue ID, and venue name, the latitude and longitude of the venue, and the foursquare venue category. We then partitioned the data into "regions" by dividing the latitude and longitude space into a grid according to the second digit of precision in the coordinates. Venues in the same grid were grouped together. In total, this procedure resulted in 55,401 regions, with 9 venues per region on average.

Topic modeling is, in the most basic sense, a model over a set of discrete observations. Although it has been applied to many other problems, the technique is most often used for text analysis, tying documents to a mixture coefficient over word distribution (so-called "topics"). Using the text analogy, in our model venue category tags are like observed words, regions like documents, and each document is characterized by as a mixture of "archetypical" neighborhoods, distributions over the venue category tags.

Notice that by modeling regions mixtures of neighborhoods we partially avoid the problem of neighborhood demarcation. We could arbitrarily drawn region boundary which happen to include multiple neighborhoods; The model fundamentally has no qualms with such melange. Moreover, we need not worry that each region contains a different number of venues. The model assumes the counts are different to begin with.

2.2 Results

We fit the foursquare venue category data to Latent Dirichlet Allocation with a fixed Dirichlet prior hyper parameter of $\alpha = 0.1$. For inference, we use the variational EM implementation introduced by [2]. We examined several different topic sizes between 10 to 50. Table 1 above shows the top 5 most likely venues for each proto-neighborhood in the 25-topic model.

Many of these topic descriptions agree nicely with our familiar neighborhood stereotypes: vineyards and wineries often find B&Bs nearby (Topic 21), Taxi Cab stations are found in the neighborhoods with pub's and theaters (Topic 7), and, indeed the video arcade is a staple of a seaside resorts community. Some of the groupings are more than obvious, such as universities (Topic 23), shopping malls (Topic 22), parks (Topic 18), airports (Topic 12), and university hospitals (Topic 1). Note that the model did not merely cluster venues of the same function together, but rather grouped them ac-

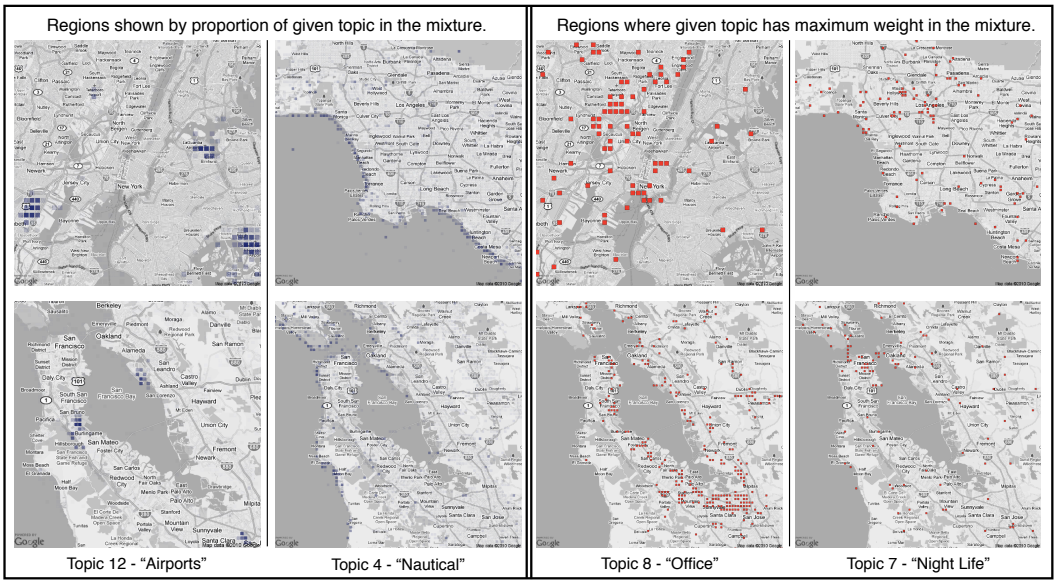

Figure 1: (Left) For each region the weight of the plots is proportional to the weight of the selected topic. We show that the nautical venues appeared on the coast line, while airport venue topic is indeed strong at known airport locations. Note that the airport regions encompass more than just the physical area of the airport venues, perhaps including venues such as hotels. (Right) Regions where the selected topic is the most dominant. Placement of these neighborhoods correspond to background knowledge of the cites.

cording to their co-occurrence patterns. Some venue types, such as pizza restaurants, appear salient in many topics, suggesting that the pizza restaurant serves an essential function in many types of neighborhoods. At the same time, we hypothesize that the pizza venues found in differing neighborhood types most likely caters to different kinds of customers, and must contain some variation beyond their categorical description that is expressed by the neighborhoods in which they reside. This presents one interesting use case for our model. Without much special training, our model could differentiate the pizzeria next to an auto garage, from the ones in a shopping mall or in an upscale residential neighborhood. In Figure 1 we plot four neighborhood topics in three different cities (New York, Los Angeles, and San Francisco). The map shows that the physical locations of the learned neighborhoods agree with our intuition about the geography of these cities.

3 Future Work

Our current plan of action is to first enlarge the venue descriptions to include free-formed text from the users. We also wish to devise extrinsic evaluations that test the utility of our neighborhood prototypes. Additionally, our current design so far ignores the spatial proximity of continuous regions or temporal trends such as gentrification and other traces of neighborhood transformation. However, we believe variations of our model could be used to account for these spatial/temporal dynamics of neighborhood formation. To the best of our knowledge, these problems have not been approached from the perspective of probabilistic generative modeling, though we feel they are apt applications for some variations of topic models [1].

We are also developing models which can address neighborhood effect studies from a topic modeling perspective, such as the modeling of neighborhoods as predictors of obesity and other diet related disease [5]. Combining disease incidence rates along side venue descriptions in a generative model may bring some interesting perspectives on often suspected correlations between the community health indicators and neighborhood effects. Several extensions to topic modeling are capable of incorporating such heterogeneous evidence in the inference [3] and we are currently examining them for this purposes.

Acknowledgments

This research was supported by NSF Trustworthy Computing grant CNS-0905562, and by Google through its support of the Worldly Knowledge Project at Carnegie Mellon University. We would like to acknowledge William Cohen, Jason Hong, Norman Sadeh and Noah Smith for their guidance on this work. We are grateful to Jacob Eisenstein for providing insightful feedback, and Aditya Lesmana for his assistance with data collection. We also thank the anonymous reviewers for their helpful comments.

References

[1] D. Blei, and J. McAuliffe. Supervised Topic Models. In *Advances in Neural Information Processing Systems 20*. (2008).

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. In *Journal of Machine Learning Research*. (2003).

[3] D. Blei, and J. Lafferty. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*. (2006).

[4] D. Dearman, and K. N. Truong. Identifying the Activities Supported by Locations Using Community-Authored Content. In *Proceedings of UBICOMP*. (2010)

[5] A. Drewnowski, C. D. Rehm, D. Solet. Disparities in obesity rates: analysis by ZIP code area. In *Soc Sci Med* **65** (12):2458-63. (2007).

[6] J. Eisenstein, B. O'Connor, N. A. Smith, and E.P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. (2010).

[7] Sadeh, N., Hong, J., Cranor, L., Fette, I., Kelley, P., Prabaker, M., and Rao, J. Understanding and capturing peoples privacy policies in a mobile social networking application. Journal of Personal and Ubiquitous Computing 13, 6. (August 2009).