Exploring venue-based city-to-city similarity measures

Daniel Preoţiuc-Pietro
Dept. of Computer Science
University of Sheffield
Sheffield, UK
daniel@dcs.shef.ac.uk

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

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

ABSTRACT

In this work we explore the use of incidentally generated social network data for the folksonomic characterization of cities by the types of amenities located within them. Using data collected about venue categories in various cities, we examine the effect of different granularities of spatial aggregation and data normalization when representing a city as a collection of its venues. We introduce three vector-based representations of a city, where aggregations of the venue categories are done within a grid structure, within the city's municipal neighborhoods, and across the city as a whole. We apply our methods to a novel dataset consisting of Foursquare venue data from 17 cities across the United States, totaling over 1 million venues. Our preliminary investigation demonstrates that different assumptions in the urban perception could lead to qualitative, yet distinctive, variations in the induced city description and categorization.

Categories and Subject Descriptors

[**Information systems**]: [Information systems applications, Spatial-temporal systems, Location based services]

Keywords

Location Based Social Networks, Foursquare, City similarity, Clustering

1. INTRODUCTION

Imagine two hypothetical cities, *Concentralia* and *Dispersia*, that are exactly the same in nearly every way, having exactly the same venues—the same universities, restaurants and parks. Suppose further that they only differ in the *spatial arrangement* of these venues, so that the venues in Dispersia are distributed uniformly throughout the city, and the venues in Concentralia are positioned more naturally—organically shaped alongside Concentralia's economic, political, and cultural evolution. One might ask, in what ways are these two cities similar or different? On the one hand, they are completely equivalent in terms of the amenities they offer—anything you can get in Concentralia you can also get in Dispersia. Yet, if you dropped a denizen of the Concentralia into the streets of Dispersia, would they perceive any resemblance to their home? In this work, we introduce some preliminary concepts for representing and comparing cities that get at the heart of the similarities and distinctions between Concentralia an Dispersia.

With the growth of smart-phones, and location-based social networks, data is being generated about human activity in urban areas at a level of detail not seen before. Numerous previous works have used machine learning methods on this social and sensor data to discover patterns within a city [1, 10]. While computational methods for understanding the landscape *within* a city are undoubtedly useful, the question of exploring relationships *between* cities using such data is relatively under-explored. There are nevertheless many practical questions in urban computing that require the *comparison* across cities. For example, a job seeker with transferable skills may wish to focus her search on a single city with jobs that best match her qualifications, rather than dispersing her search efforts across multiple cities. Likewise, a large corporation looking to expand its locations might perhaps select cities it wishes to expand into before considering particular sites or neighborhoods. Additionally, many within-city computations might be aided by modelling a city's relationship to other cities. For example, a person buying or renting a home in a new city might want to be able to compare the neighborhoods of the city to other neighborhoods in different cities.

One issue in comparing spatial regions such as cities is the normalization of absolute data, since often raw data from two different contexts are incomparable. Another challenge is the question of how to account for spatial effects. In our hypothetical comparison, the only difference between Concentralia and Dispersia was in the spatial distribution of their venues. In practice, often spatial effects can be accounted for by partitioning the data into discrete spatial units of aggregation, e.g. grids or neighborhoods, and then ignoring any spatial dependencies in the data within these discrete units [2, 6]. Determining the proper granularity of discretization, however remains a difficult problem.

In this preliminary investigation, we explore different of units of spatial aggregation, namely grids, neighborhoods, and the city as a whole, in the representation of cities as vectors over venue types. In highlighting the differences between these approaches, we hope to encourage future researchers to re-

UrbComp 13, August 11–14, 2013, Chicago, Illinois, USA.

alize that the level of spatial aggregation is an important factor when featurizing a city in terms of its venues.

2. RELATED WORK

In recent years the influx of location-aware social and sensor data has inspired a score of new studies. In Eisenstein et al. [3] and Wing and Baldridge [8], the authors use social media to examine lexical diversity across cities in the United States, applying their algorithms to automatically predict the location of microblog messages. Zheng et al. explored how sensor equipped taxi-cabs could be used to identify traffic engineering flaws that could potentially lead to high congestion [10]. Cranshaw et al. use Foursquare check-ins to redefine the notion of a neighborhood by clustering city venues into contiguous areas reflecting the check-in patterns of like-minded people [1]. A number of works have also looked into using social data to model semantic qualities of a place. In Cranshaw and Yano [2] and Noulas et al. [6], the authors partition cities into grids, and then cluster the resulting grids according the types of venues found there, revealing patterns in land usage across cities. Yuan et al. use properties of visitation patterns a region and other spatially embedded meta-data to try to discern the region's functional category [9]. Hong et al. [4] attempted to induce the set of geographically biased topics, with application to the topic popularity detection. Preoţiuc-Pietro and Cohn [7] use the semantic information of venues to group users based on their behaviour and predict their future movements.

3. CITIES AS COLLECTIONS OF VENUES

We hypothesize that one natural way to characterize a city is by the ensemble of amenities it offers. For example by tallying the amount of parks, bars, or universities it has relative to all other types of venues, one can get a sense for what a city is like. Although collecting such detailed data about the places in a city was in the past challenging, location-based social systems such as Foursquare often prompt their users to tag the locations they check-in to with descriptive categorical labels. Exploiting such incidental, geo-identified semantic information from social media users could be one promising approach to urban folksonomy.

In this preliminary investigation, we explore three different vector representations of a city as normalized counts of the venue categories seen there. Each representation scheme uses the same raw data, the counts of venue category tags. They differ only in how these counts are aggregated and normalized. They all share the same basic idea of cities as ensembles of amenities, but each reflect slightly different ideas about the unit of special aggregation in urban perception.

First we define the *bag of venues* representation of a spatial region r. Suppose there are n venues located within r, and each is chosen from a global set of m venue categories, then we define the m-dimensional bag of venues vector $c(r) = \frac{1}{n}(c_1, c_2, \ldots, c_m)$, where c_i is the number of venues of category i within r, and n is the total number of all the venues ($n = \sum_{i=0}^{m} c_i$).

City-centric representation: In this representation we do no spatial aggregation within the city. Viewing its venues as spatially exchangeable, we representing a city x as its bag of venues vector, $c(x)$. Under this representation, Concentralia and Dispersia are indistinguishable.

City Abbreveations		
Atl - Atlanta	Aus - Austin	Bal - Baltimore
Bos - Boston	Chi - Chicago	Col - Columbus
Hou - Houston	LA - Los Angeles	NY - New York
Phi - Philadelphia	Phx - Phoenix	Pit - Pittsburgh
SA - San Antonio	SD - San Diego	Sea - Seattle
SF - San Francisco	Was - Washington, D.C.	

Table 1: City abbreviations used in the article.

Grid-centric representation: In this representation, we first divide the area of city x into a set of k of equally sized grid regions $r_1, r_2, \ldots, r_k$ (in our experiments we use grids of 0.01 units latitude by 0.01 units of longitude), and we aggregate across these grids. First computing $y = \sum_{i=1}^{k} c(r_i)$, we then normalize the result, to represent the city x as $y/|y|_1$. Note that because each $c(r_i)$ is itself a normalized vector of counts, each region contributes equally to the final vector. This representation is thus better at measuring how certain categories concentrate within the regions r_i.

Neighborhood-centric representation: This representation is equivalent to the grid-centric approach, but here, rather than aggregating over arbitrarily chosen grids, we aggregate over municipal neighborhood boundaries, hypothesizing that neighborhoods most effectively communicate the city's character.

4. DATA

Exploring these ideas empirically using real world data requires that we gather both the description, or the categorizations, of the venues in a city, as well as information about the city's municiple neighborhood boundaries. For the venues, we collected data from the widely used location-based Social Network (LBSN) Foursquare. Users of Foursquare "check-in" to their current location on their mobile device by selecting it from a list of nearby named venues. Their check-in is then broadcast to their social connections. Foursquare users can also specify a hierarchical categorical description to a venue, such as "Restaurant" and "Mexican Restaurant". We call higher-level (more general) categories the *primary category*, and we call lower level (more specific) categories the *secondary category*. An interesting side-benefit of such collaborative tagging is that as the system's user base increases, an accurate and up-to-date crowd-sourced representation of the venues types within a city is naturally accumulated. In this work we exploit this natural data for our empirical study. The venue information is easily accessible through a public API, and all venues are annotated with categories of different granularities which represent a natural semantic grouping for venues.

We collected Foursquare venues from 17 cities across the United States, selecting the largest ten cities by municipal population, and seven additional major cities from diverse geographic areas of the country. For each of these cities, we collected boundary files from the online real estate company Zillow[1] that detail the city's municipal neighborhood borders. Given these boundaries, we determined in which neighborhood each Foursquare venue is located.

[1] `http://www.zillow.com/howto/api/neighborhood-boundaries.htm`

City	C-s and N-s	C-s and G-s	N-s and G-s
Phx	0.55	0.50	0.79
Chi	**0.38**	**0.36**	0.69
LA	0.58	0.51	0.75
Bal	0.60	0.54	0.82
Atl	0.50	0.58	0.58
Aus	0.60	0.64	0.69
Was	0.48	0.50	0.63
Col	0.58	0.60	0.80
SA	0.66	0.66	0.79
Pit	0.72	0.67	0.86
Phi	0.57	0.58	0.69
Hou	0.64	0.54	0.72
Sea	0.55	0.47	0.55
SD	0.72	0.64	0.89
SF	**0.45**	**0.45**	0.67
Bos	0.47	0.50	0.47
NY	**0.27**	**0.26**	0.63

Table 2: Kendall Tau correlations ($-1 \leq \tau \leq 1$) between representation similarity rankings for each city. All correlations except those in bold are signficant at p=0.05.

Venue data was collected in September 2011 using the Venues API[2] of Foursquare. First venues were crawled within a 40 mile radius of the city's center. We discarded any venues that either fell outside of the city boundary, or that could not be mapped to a neighborhood of the city (even if it they fell inside the city). This ensured that each representation method is applied to the same set of data points. Discretization of venues into categories is performed naturally by using either the principal or secondary categories of the venues depending on which analysis we perform. There are 9 principal venue categories and 259 secondary venue categories. In total, once this processes was complete, our final data set consisted of 1,130,621 venues across the 17 cities, with the best represented city, New York, having 210,335 venues and the least represented city, Pittsburgh, having 19,830 venues.

5. COMPARING CITIES

In this section we explore the effects of the different representation methods. Using the data described above, we created the vector representations of each of the 17 cities using the three methodologies discussed in Section 3, for both primary and secondary venue category tags, resulting in a total of 6 different representation. In our results we denote them by **C-p**, **C-s**, **N-p**, **N-s**, **G-p**, **G-s**, where we **C**, **N**, and **G** denote the city, neighborhood, and grid-centric representations respectively, and **p** and **s** denote the use of primary and secondary categories for features. Our hypothesis is that, depending on how we aggregate and normalize the venue description vector, we will end up with much different representations of the cities, which will lead to differing results when applied to various analytic tasks. In the next few sections we show these representations results in discernibly different outcomes.

5.1 City similarities

One straightforward way to see the effects of the different vector representation is to conduct similarity analysis. Representing each city as a point in a vector space, we can compute the pairwise similarities between the cities using the standard cosine similairty[5][3]. Then for each city, we ranked all other cities by their similarity to the said city. We repeated this ranking for all vector representations. We then quantify the difference among the computed rankings for each city using the Kendall Tau rank correlation coefficient. Results are presented in Table 2. Although most representation rankings are significantly correlated, results show a wide range in the coefficients across the cities, indicating the importance of the representation. In general, the largest differences in representation were seen between the city centric representation and the other two. Similar results were seen when primary category counts were used as features.

New York			Chicago			Pittsburgh		
C-s	**N-s**	**G-s**	**C-s**	**N-s**	**G-s**	**C-s**	**N-s**	**G-s**
Sea	Sea	Sea	LA	Phi	Phi	Bal	Phi	Bal
SF	SF	Phi	Phx	Bal	Bal	Phi	Bal	Phi
Bos	Was	Was	Phi	LA	LA	Col	SD	Col
LA	Bal	SF	SD	Atl	SF	SD	Col	SD
Chi	Col	Pit	Bos	Pit	Bos	Aus	Atl	LA

Table 3: Top 5 most similar cities for New York, Chicago, and Pittsburgh under the three representations.

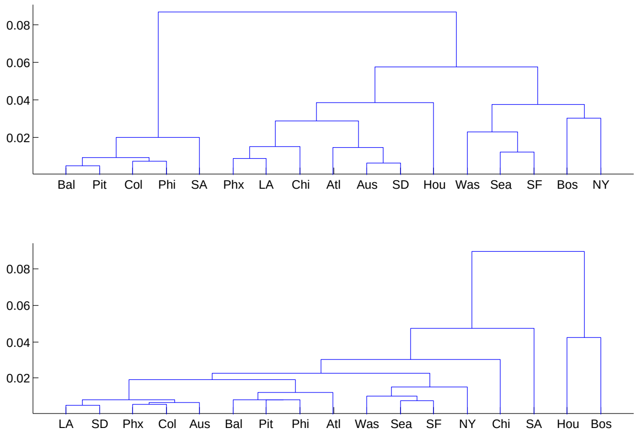

Figure 1: Complete linkage heirarchical clustering using C-s (top) and N-s (bottom).

Given these similarities between cities, one query a person may like to make is what cities are most similar. We present for a few selected cities the top most similar cities as obtained by the three methods using the secondary categories. The results, presented in Table 3, indicate that the chosen representation can have significant impacts on such queries.

5.2 Clustering Cities

To visualize how the cities relate to one another under each representation, we using agglomerative hierarchical clustering with complete linkage to cluster cities into groups [5]. Al-

[2] `https://developer.foursquare.com/overview/venues`

[3] Cosine similarity of course is not the only applicable similarity measurement, though it is a natural choice when comparing counts.

Venue category	N-p/C-p	G-p/C-p	G-p/N-p
Arts	-0.17	-0.23	-0.06
Transport	-0.09	-0.10	-0.00
Shops	-0.03	-0.04	-0.00
Food	-0.13	-0.19	-0.05
Parks	0.24	0.27	0.02
Nightlife	-0.17	-0.24	-0.07
Residence	0.20	0.28	0.06
College	-0.10	-0.11	-0.01
Professional	-0.01	-0.01	0.00

Table 4: Relative change between values of different representations

though other options exist, we prefer hierarchical clustering for its ease in interpretation of the results as a dendrogram, and since its not intuitive to select the number of clusters a priori. Figure 1 shows the the dendrogram structures for **C-s** and **N-s**. Results indicate that, although there are some similarities observed in the two groupings (e.g. Sea-SF, Bal-Pit-Phi), they are qualitatively distinct. One intuitive presumption would be that geographic distance between cities would play a large role in the clustering. Although we did observe some natural geographic clusters in the dendrograms (e.g. LA and SD; Phi, Pit and Bal), when examining the Kendall Tau rank correlation coefficient between our venue based similarity rankings and the rankings formed with great circle distances between the city centers, we found no significant correlation at a level of $p = 0.05$.

6. QUALITATIVE ANALYSIS

To understand our results further, we examine which categories get over or under represented in each aggregation scheme. To see this, for each (principal) category, we compute the difference of the values of two representations of the same city divided by the value of the first representation to measure relative change. In Table 4 we show the average of this value across all the cities. If this value is negative, then the This shows that some venue categories, such as 'Parks' or 'Residence' are under represented by **C-p**, while others such as 'Nightlife', 'Arts' and 'Transport' are over represented. We suppose that this happens because the venues in the former category tend to cluster, forming distinctive spatial units within the cities. The venues from the later category are either over represented in the data or they are 'spread' in all the neighborhoods or grids, with no significant spatial unit having predominantly this type of venues. This can be the case with 'Nightlife' spots, which, although usually are spatially close, they are not the predominant category, usually being joined by Shops or food outlets. We also notice that the grid representation is more 'aggressive', producing larger relative differences. This is somewhat intuitive; For example, in the 'Residence' case we can expect that predominantly residential suburbs that belong to one neighborhood probably span multiple grids, while grids that include 'Arts' and 'Nightlife' venues are probably downtown, where spatial distances are smaller, resulting in more densely populated grids. Moreover, the neighborhoods are expected to mimic the underlying 'significance' of groups of venues in a city.

As examples, we look at the highest relative differences in the Neighborhood and City representations. New York has a large absolute value of the ratio for 'Parks; (0.61 increase) and for 'Arts' (-0.26 decrease) meaning that while parks are concentrated in their own neighborhoods (e.g. Central Park consisting mostly of park related venues) and arts spots are dispersed across the city. San Francisco also stands out for having the largest (0.79) increase in 'Parks,' very high increase in the 'Residence' category (0.31) and the highest decrease in 'Professional' venues (-0.19), closely followed by New York and Washington. This show that these last 3 cities have professional buildings and offices spread around the entire city more than the others.

7. CONCLUSION

In this preliminary investigation we have presented different methods for comparing cities as vectors of venue categories. We have identified and emphasized the choice of aggregation level and shown that it can have significant quantitative and qualitative effects for city-to-city comparisons. We have presented the results of city similarities as well as an analysis on the frequencies of different venue categories and how each representation may affect them. In future work, we want to carry on a similar analysis between neighborhoods of cities, in order to identify similar neighborhoods across cities and to get a better understanding of cities as collections of individual neighborhoods. Finally, we hope that our work will motivate future studies into how to best characterize a city in terms of its venues.

8. REFERENCES

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

[2] J. Cranshaw and T. Yano. Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling. In *NIPS, Workshop of Computational Social Science and the Wisdom of the Crowds*, 2010.

[3] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, pages 1277–1287, 2010.

[4] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, pages 769–778, 2012.

[5] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval.* Cambridge University Press, 2008.

[6] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *ICWSM 2011*, 2011.

[7] D. Preoţiuc-Pietro and T. Cohn. Mining User Behaviours: A Study of Check-in Patterns in Location Based Social Networks. In *Web Science*, 2013.

[8] B. P. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *ACL HLT*, volume 1, pages 955–964, 2011.

[9] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *ACM SIGKDD*, pages 186–194, 2012.

[10] Y. Zheng, Y. Liu, J. Yuan, and X. Xie. Urban computing with taxicabs. In *UbiComp*, 2011.