"What Is the City but the People?"* Exploring Urban Activity Using Social Web Traces

Emre Çelikten
Computer Science
Aalto University
Helsinki, Finland
emre.celikten@aalto.fi

Géraud Le Falher
Inria, Univ. Lille
CNRS UMR 9189 – CRIStAL
F-59000 Lille, France
geraud.le-falher@inria.fr

Michael Mathioudakis
Helsinki Institute for
Information Technology
Aalto University
Helsinki, Finland
michael.mathioudakis@hiit.fi

ABSTRACT

We demonstrate GeoTopics, a system to explore geographical patterns of urban activity. The system collects publicly shared check-ins generated by `Foursquare` users, that reveal who spends time where, when, and on what type of activity. It then employs sparse probabilistic modeling techniques to learn associations between different regions of a city and multi-feature descriptions of urban activity. Through a web interface, users of the system can select a city of interest and explore visualizations that highlight how different types of activity are spatially and temporally distributed in the city.

We discuss the opportunities that web data offer to understand urban activity and the challenges one faces in that task. We then describe our approach and the architecture of GeoTopics. Finally, we lay out the demonstration scenario.

Keywords

Urban Computing; Location-Based Social Networks; Probabilistic Models

1. INTRODUCTION

Modern cities are massive, dynamic and complex systems. They buzz with activity, typically taking place at venues such as restaurants, shopping malls, parks, and so on. Many aspects of a citizen's life depend on how this activity is distributed across a city – for example, where to look for an apartment, where to go spend a frenetic Friday evening, or how much to price a house. Obtaining insights about the geographical structure of urban activity is thus a task of high potential impact, especially as an ever-increasing number of people live in cities [1]. In this context, the growing amount of data produced by urban dwellers on the social web offer new opportunities for analysis supporting that goal.

The system we present makes use of `Foursquare` data, a popular location-based social network. One of the main functionalities of the platform is to enable its users to generate *check-ins* that inform their friends of their whereabouts. Each check-in contains information that reveals *who* (which user) spends time *where* (at what location), *when* (what time of day, what day of week), and doing *what* type of activity (as induced by the kind of venue that hosts the activity, e.g. dining at a restaurant or shopping at a grocery store).

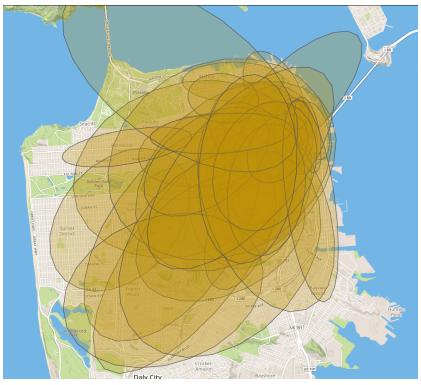


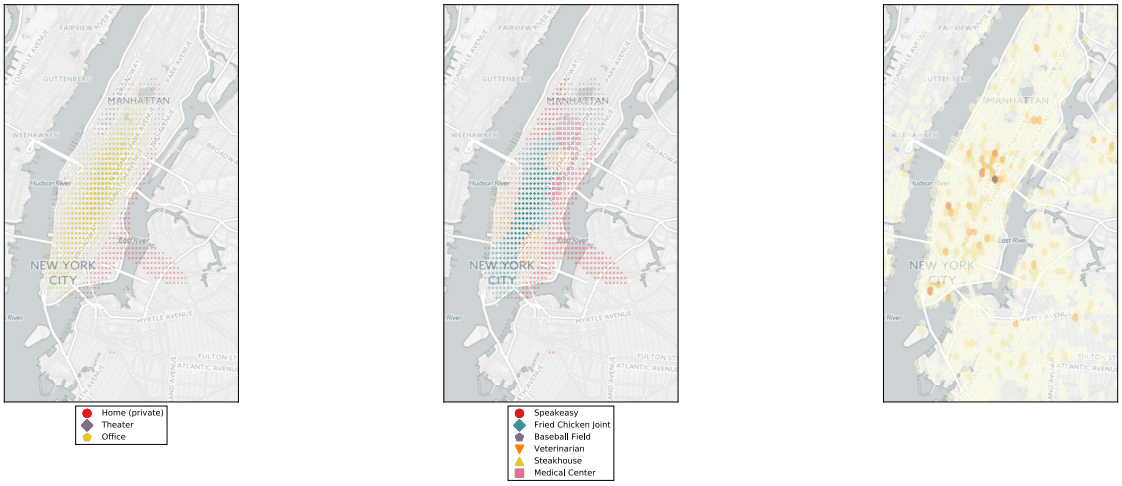
Figure 1: The 59 regions discovered in San Francisco. (Maps provided by ©OpenStreetMap [7]).

The system collects `Foursquare` data automatically and processes them to learn whether and how different regions in a city are associated with different types of activity. In particular, the system employs sparse probabilistic modeling techniques to learn a decomposition of a city into a small, optimal number of possibly overlapping regions, each associated with a different description of activity. As an example, one might discover that the south of a city is strongly associated with restaurants, bars, and night-clubs that are active in the evening, while the north of the city is associated with cafeterias and parks that are active in the morning of weekends. Such a decomposition is an easily interpretable way to study the distribution of different types of activity across a city.

Furthermore, the employed modeling techniques allow us to answer more elaborate questions about urban activity. Specifically, they offer a principled and accurate way to quantify the degree to which any point on the city map is associated with any type of activity. One can use this information to plot heatmaps that demonstrate how a particular type of activity is distributed across a city, but also what the most distinctive types of activity across different city locations are. As an example, we invite the reader to in-

*William Shakespeare, *Coriolanus*, Act III, Scene I.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2901922.



(a) Frequent categories. (b) Distinctive categories. (c) Distribution of restaurants.

Figure 2: Feature heatmaps produced by our GeoTopics system in Manhattan. (Maps provided by ©OpenStreetMap [7]).

spect: Figure 2c, which demonstrates how **restaurants** are *distributed* across Manhattan; Figure 2a, which demonstrates what venue category is most *frequent*; and Figure 2b, which demonstrates what venue category is most *distinctive* for different locations across the same region (i.e., it appears at a given location much more often than at other locations, even if it is not the most frequent there).

Related Work Urban Computing is an emerging field of research that aims to extracts valuable insights through computationally assisted data analysis [11]. In the vicinity of our work, finding cohesive geographical regions within cities has been approached using a variety of data sources, such as cellphone activity [8], geotagged tweets [5], social interactions [6], types of buildings [3], or public transport and taxi trajectories [9]. In that context, Location Based Social Networks (*LBSNs*) have also proven a rich source of data and utilized by recent works. The Livehoods project [4], for example, processes `Foursquare` checkins and makes use of spectral clustering to discover clusters of nearby venues with high overlap of visitors. Also using `Foursquare` data, [10] describes venues by category, peak time activity and a binary touristic indicator. Venues are clustered in hotspots along all these dimensions by the OPTICS algorithm. Finally, [2] builds a platform that aggregate urban data from various sources including LBSN and visualize spatio-temporal pattern of activity, although without defining dynamically their spatial extent as in our work. Unlike earlier work, the probabilistic engine in the backend of our system allows us to learn and represent information at much finer level of detail.

Outline In the rest of this proposal, we provide further description of the different components of the system. Subsequently, we describe its architecture and lay out the demonstration scenario.

2. ANALYZING URBAN ACTIVITY

2.1 Data

Our data consists of geo-tagged activity from `Foursquare`, a popular location-based social network that, as of December 2015, claims more than 50 million users[1]. It enables users to share their current location with friends, rate and review venues they visit, and read other users' reviews. `Foursquare` users share their activity by generating *check-ins* using a dedicated mobile application[2]. Each check-in is associated with a web page that contains information about the user, the venue, and other details of the visit. Each venue is also associated with a public web page that contains information about it — notably the city it belongs to, its geographic coordinates and the corresponding category, such as *Food* or *Nightlife Spot*.

According to `Foursquare`'s policy, check-ins are private by default, i.e., they become publicly accessible only at the users' initiative. This is the case, for example, when users opt to share their check-ins publicly via `Twitter`[3], a popular micro-blogging platform. Therefore, to collect publicly shared check-ins, the system filters tweets containing a specific URL pattern denoting a check-in and subsequently queries `Foursquare`'s API to obtain full information about each check-in.

To analyze how activity is distributed across a city, the system takes a venue-centric view of the data. Specifically, it associates the following information with each venue.

- A geographic `location`, expressed as a longitude-latitude pair of geographic coordinates.
- The `category` of the venue, as specified by `Foursquare`'s taxonomy (e.g., 'Art Gallery', 'Irani Cafe', 'Mini Golf'). If more than one categories are associated with one venue, then it keeps the one that is designated as the 'main category'.
- A list of all check-ins associated with this venue in the working dataset. Each check-in is a triplet that contains the following data: (i) the unique identifier of the `user` who performed it; (ii) the `day of the week` when the check-in occurred, expressed as a categorical variable with values *Monday*, *Tuesday*, ..., *Sunday*; (iii) the `time of the day` when the check-in occurred, expressed as a categorical variable with values *morning*, *noon*, ..., *late night*.

[1] `https://foursquare.com/about/`.
[2] The Swarm application, `http://www.swarmapp.com`.
[3] `http://twitter.com`

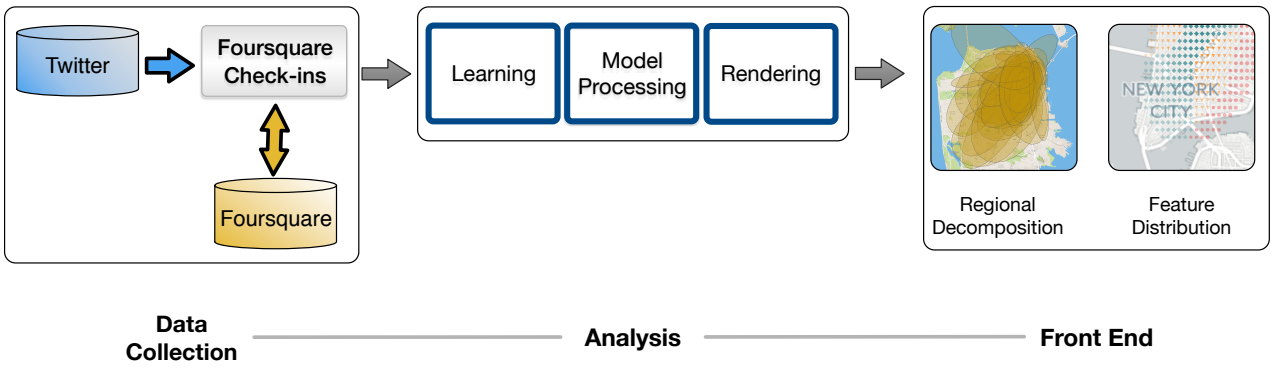


Figure 3: Organization of our three independent modules into the GeoTopics system described in the main text. (Maps provided by ©OpenStreetMap [7]).

2.2 Modeling

Given the venues within a city and all associated collected activity (venue location, venue category, check-in user, and check-in time), the system employs topic modeling inspired unsupervised learning techniques to produce a decomposition of the city into – possibly overlapping – regions[4]. Each region is associated with its own probabilistic description of the features described earlier. As an example, one topic might generate venues (data points) that are located in the south of a city (feature: `location`) and are particularly popular in the morning (feature: `time of the day`), while another might generate venues that are located in the north of a city (feature: `location`) and predominantly restaurants, bars, and night-clubs (feature: `category`).

The approach we take is automatic, in the sense that it does not require any manual parameters tuning or arbitrary strong assumptions about the data. For instance, it does not require us to select the optimal number of regions in advance or specify the granularity of the analysis. In the system implementation, the same software is applied to produce the decomposition for each city in the data.

2.3 Exploring

The discovered regions are themselves of immediate interest, as they provide an interpretable way of exploring the distribution of different types of activity in the city. As an example, Figure 1 shows how activity in San Francisco is decomposed in a small number of overlapping regions. For each such region, the system has learned a probabilistic description of the types of activity encountered therein. Namely, what fraction of the total activity is associated with venues of different categories, times of the day, and days of the week.

Furthermore, the system uses the learned model instances to answer more elaborate questions about the data. What is the most frequent venue category at given location? What is the most distinctive category at a given location? How are venues of a particular category distributed across the city? The system processes the modeling results and constructs maps of the city annotated to provide answers to those questions. Likewise, similar questions are asked for other features in the data, such as the time and day of check-ins – and the system produces corresponding maps.

[4]Details of the approach are described in a manuscript that is currently under review [12].

3. ARCHITECTURE

The architecture of the system is shown in Figure 3. As depicted there, it consists of three mostly independent modules, namely *data collection*, *analysis*, and the *front-end*.

Data collection The two main tasks of this module are to (i) filter the `Twitter` *gardenhose* sample to look for publicly shared check-ins, and (ii) subsequently query `Foursquare` in order to obtain data associated with each check-in. The module also performs data-cleaning and filtering tasks. Its output is a database of venues and their associated activity for a period of time. On one hand, this module stands alone, as such data can be used for other kind of analysis. On the other hand, our flexible model of urban activity allows other datasources to contribute to our database and later provide richer analysis.

Analysis The two main tasks of this module are to (i) fit a model on the data for each city, and (ii) process the resulting models and produce the maps used for exploration by the next module. Model training takes a few hours on 8 cores using the `numpy` Python library for data acquired over the period of one month.

Front-end The system provides a user interface to enable the exploration of analysis results. The user is able to select a city of interest and a time period (e.g., January 2016) and explore the analysis results produced for that city by the two previous modules.

4. DEMONSTRATION

The system is demonstrated through its publicly accessibly webpage. The users will interact with it by exploring the following concepts.

Regional decomposition The user selects one city and one period of interest (e.g., 'San Francisco, January 2016') to obtain a decomposition of the city into a small number of regions, each associated with a description of activity. The users will be able to select each region in the aforementioned decomposition to obtain its description. The description is provided in the form of relative frequencies, that describe how often different types of activity appear in that region (e.g., '15% of activity appears at restaurants') and how more often compared to the entire city (e.g., '2 times more frequently than the rest of the city').

Frequent and distinctive feature values The user selects one feature of the data, e.g., the category of venues or the

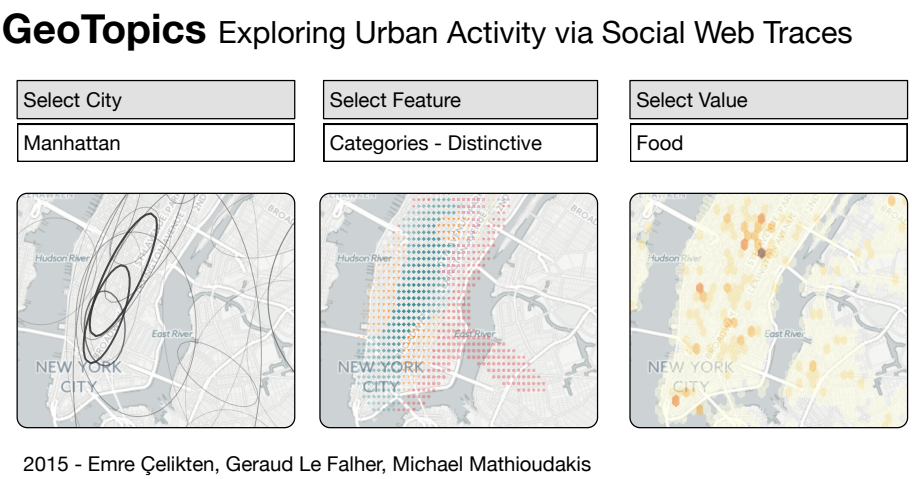


Figure 4: Front-end visualization examples. (Maps provided by ©OpenStreetMap [7]).

time of the day of check-ins, and obtains two heatmaps: one that shows which feature value is most frequent at each point in the city; and one that shows which feature value is most distinctive at each point in the city (in the sense that it appears much more frequently than in the rest of the city). For example, by inspecting the heatmaps, one might discover that *restaurants* is the most common category type near Times Square in Manhattan, and *theaters* is the most distinctive category type for the same area.

Geographic distribution of feature values Finally, the user picks one value for the feature selected in the previous step (e.g., value 'restaurant' for feature *venue category*, or value 'morning' for feature *check-in time of the day*). The system then provides a distribution of activity associated with the selected feature value across the city.

Example visualizations of the system are shown in Figure 4. The front end of the system will be made publicly available before the demonstration, together with code for the various modules of the system[5]. Maps are credited to ©OpenStreetMap contributors [7], for more information see http://www.openstreetmap.org/copyright.

5. ACKNOWLEDGEMENTS

This work is supported by the European Community's H2020 Program under the scheme 'INFRAIA-1-2014-2015: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem'.

6. REFERENCES

[1] World Urbanization Prospects, the 2014 Revision: Highlights. Technical report, United Nations, Department of Economic and Social Affairs, New-York, 2014.

[2] S. Bocconi, A. Bozzon, A. Psyllidis, C. Titos Bolivar, and G.-J. Houben. Social glass: A platform for urban analytics and decision-making through heterogeneous social data. In *Proceedings of the 24th International Conference on World Wide Web*, pages 175–178, 2015.

[5] `http://mmathioudakis.github.io/geotopics/`

[3] Z. Cao, S. Wang, G. Forestier, A. Puissant, and C. F. Eick. Analyzing the Composition of Cities Using Spatial Clustering. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, pages 14:1–14:8, New York, NY, USA, 2013. ACM.

[4] J. Cranshaw, J. I. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM*, pages 58–65, 2012.

[5] V. Frias-Martinez and E. Frias-Martinez. Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence*, 35:237–245, 2014.

[6] J. R. Hipp, R. W. Faris, and A. Boessen. Measuring 'neighborhood': Constructing network neighborhoods. *Social Networks*, 34(1):128–140, 2012.

[7] OpenStreetMap. Copyright and license, http://www.openstreetmap.org/copyright.

[8] J. L. Toole, M. Ulm, M. C. González, and D. Bauer. Inferring Land Use from Mobile Phone Activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, pages 1–8, New York, NY, USA, 2012. ACM.

[9] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong. Discovering Urban Functional Zones Using Latent Activity Trajectories. *Knowledge and Data Engineering, IEEE Transactions on*, 27(3):712–725, 2015.

[10] A. X. Zhang, A. Noulas, S. Scellato, and C. Mascolo. Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. In *ASE/IEEE SocialCom*, pages 69–74, Sept. 2013.

[11] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban Computing: Concepts, Methodologies, and Applications. *ACM Transaction on Intelligent Systems and Technology*, 5(3):38:1–38:55, Sept. 2014.

[12] E. Çelikten, G. L. Falher, and M. Mathioudakis. Modeling urban behavior by mining geotagged social data. *TBD - under review*, 2015.