Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Scalable Urban Data Collection from the Web

Rijurekha Sen
MPI-SWS
Saarbruecken, Germany
rijurekha@mpi-sws.org

Daniele Quercia
Bell Laboratories
Cambridge, UK
dquercia@acm.org

Carmen Vaca Ruiz
Escuela Superior
Politecnica del Litoral
cvaca@fiec.espol.edu.ec

Krishna P. Gummadi
MPI-SWS
Saarbruecken, Germany
gummadi@mpi-sws.org

Abstract

Easy access to different necessities of daily life makes a city more livable. This has motivated urban planning researchers to quantify urban accessibility from official city data. However, due to the manual nature of data collection, these earlier survey based analyses were limited in scope and scalability, and mostly offered insights on cities of developed countries like the UK and the USA.

Using Google Places data that is crowd-sourced around the world, this paper gathers walkability information for twenty-five cities across five continents. We detail the collection methodology of this unprecedented dataset and show useful applications of this data in urban analysis: e.g., how different areas within a city compare against each other in terms of accessibility and which areas in a city would benefit the most from the least intervention.

1 Introduction

The growing demand for walkable neighborhoods has made services that calculate walkability (e.g., walkonomics.com, walkscore.com) popular among real estate agents, health-care agencies, and environmentalists. However, these sites needed to process and gather a variety of datasets, which can be financially prohibitive (Quercia et al. 2015). In comparison with these prior works on quantifying accessibility, we propose a scalable method using Google Maps public APIs to crawl web data. This scalable and fine-grained data collection methodology enables us to measure accessibility not only for different areas in a particular city, but for different cities in the world.

Similar to our approach, (Cranshaw et al. 2012) and (Vaca et al. 2015) use web data to identify functional uses in a city. They use data from location based social network Foursquare. However, Foursquare data is sparse for many cities, especially in developing countries. Instead, we leverage the wider coverage of Google Maps data, which is crowd-sourced in almost all cities in the world.

Our data collection methodology based on Google Maps API is detailed in Section 2. An illustrative analysis using this fine-grained dataset for recommending urban interventions is discussed in Section 3. Directions of future explorations are outlined in Section 4 and we conclude the paper in Section 5.

2 Urban Web Data

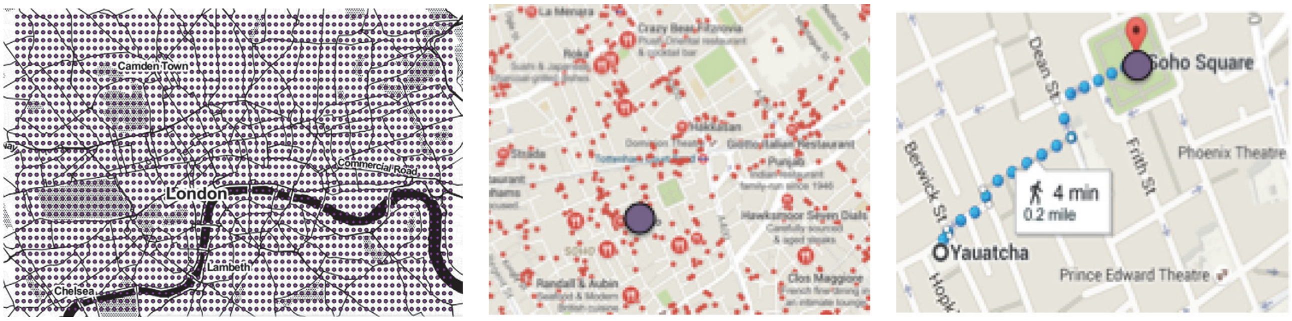
To determine what is accessible where, we need to measure the walking distances between an area and different daily life facilities. We opt for the Google Maps public API, mainly because it is widely available around the world. We propose a crawling methodology that is *reproducible* (others can repeat it) and *scalable* (the collection of data for a variety of cities does not require a prohibitive number of API calls).

Data collection method: Our data collection method is illustrated in Figure. 1. We divide each city into $200m$ X $200m$ square grids, and take the centre of each such square as our *centroid* or *area* for analysis. The <lat, lon> coordinates of these centroids or areas are input to the Google Places API. The outputs of the Places API are the details of places in different categories (described later in Table 1), nearby to the area under consideration.

Once a list of places is obtained for each area for the different categories, the nearest place in each category is taken. The <lat, lon> coordinates of the area and the place nearest to that area, are then input to the Google Distance Matrix API. The outputs of the Distance Matrix API are the walking distances and times, to travel from the area to the nearest place. We obtain these values for the nearest place in every category, for each area in the city.

Google does not currently include real time traffic and other such information in its travel time results. Thus the time values are static information, simply based on distances and assuming a typical walking speed. In our subsequent analyses, we therefore mostly use the walking distances and design our metrics and methodologies based on them.

Categories used: The Google Places API offers detailed information in different place categories. The categories used in our analysis are given in Table 1, along with their



(a) 200x200 m grids in London, centers of which are used for Google Places API queries

(b) Sample Google Places API query for category *restaurant* from a center at Soho, London

(c) Sample Google Distance API query to nearest restaurant from the center at Soho, London

Figure 1: The three main steps necessary to obtain our Urban Web Data

Categories	Purposes
bar—restaurant, bakery, cafe, convenience store—grocery or supermarket	food and daily necessities
bus station, taxi stand, train station— subway station, bicycle store, parking, gas station	transportation
shopping mall—department store, clothing store—shoe store—jewelry store	shopping and retail
doctor—dentist, hospital, beauty salon—hair care—spa—gym	health services
atm—bank	financial services
school—university	education
art gallery—museum, book store, library, movie rental, movie theater, night club	entertainment and tourism
stadium, amusement park—rv park— campground—zoo—aquarium, park	sports and outdoor activities
fire station, police	safety
church—hindu temple—mosque— place of worship—synagogue	religion

Table 1: List of facility categories.

common purposes in urban lives. To reduce the number of API calls and remain within the API query limits imposed by Google, we combine some very similar categories together using the '|' operator. There are 30 different category blocks, after the '|' based combination. Thus for each centroid or area, there are 30 calls issued to Google Places API, to get the nearby places in those 30 categories.

Cities crawled: We repeat those three steps for as many as 25 cities in both developed and developing countries across the five continents (Table 2). They either belong to the developed or industrialized countries, mostly in Europe, North America and in some countries of East Asia. Or they belong to developing countries in South Asia, Africa or South America.

Cities	Characteristics
Barcelona, Berlin, London Milan, Paris, Rome	Industrialized; Europe.
Chicago, New York, San Francisco Seattle, Toronto, Washington	Industrialized; North America.
Beijing, Singapore, Tokyo	Industrialized; Asia.
Bengaluru, Buenos Aires, Delhi, Jakarta, Kuala Lampur, Mexico, Moscow, Mumbai, Nairobi, Rio	Developing; India, South America, Africa.

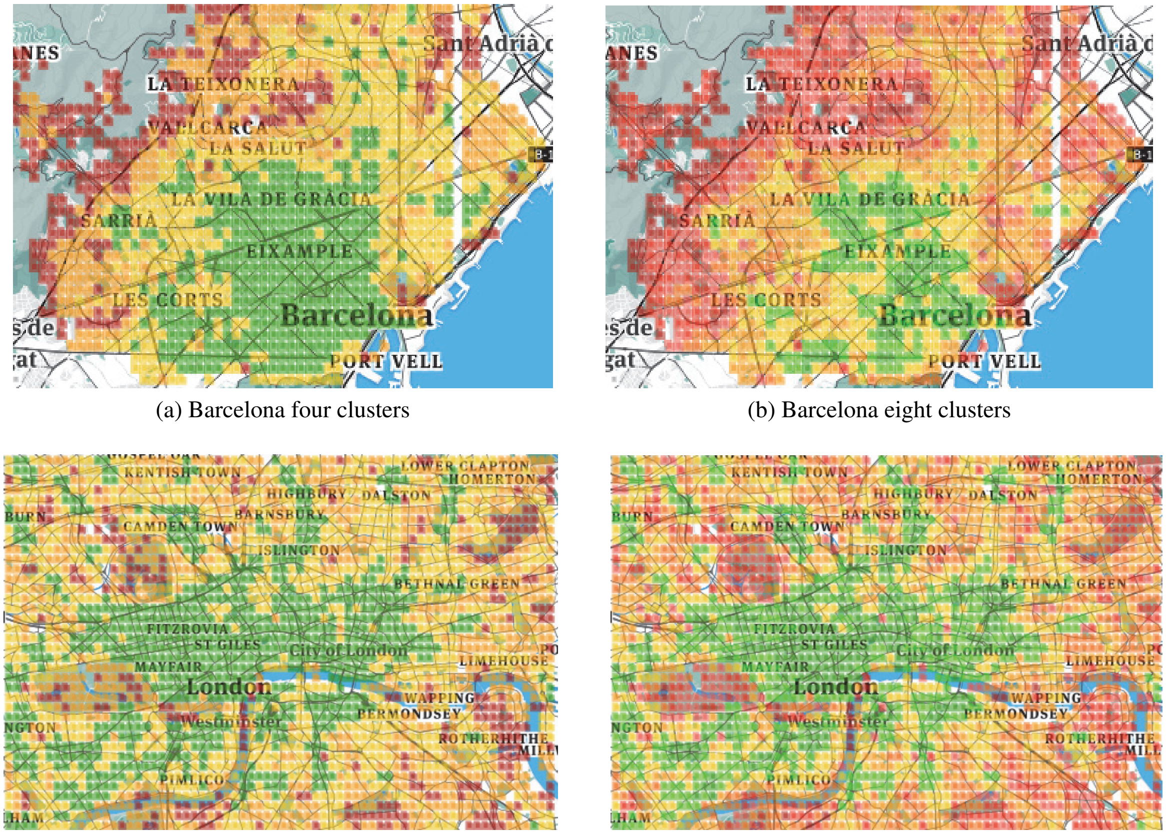
Table 2: Cities under study.

Area-by-category distance matrix: We represent our crawled data in an area-by-category distance matrix. The rows represent the areas in the city, which are the centers of the $200m$ X $200m$ squares, into which the city is divided. The columns are the 30 categories described in Table 1. The value for each area-category cell is the distance between that area and the nearest venue in that category, from that area. We construct one such matrix for each city and all our subsequent analyses will be based on these matrices.

3 Urban Interventions

Having our data organized as area-by-category distance matrices, we now demsonstrate the utility of this fine-grained dataset to analyze urban accessibility and inform simple interventions. To this end, we need to determine which areas are rich (in terms of accessibility) and which are poor.

We cluster the areas in a city that are similar to each other in terms of accessibility diversity. Therefore we cluster the areas based on how diverse are the facilities which are within walking distance of a particular area. We run K means clustering and the resulting clusters for $K = 4$ and 8 are given in Figure 2 for Barcelona and London. Red denotes lower feature values or less diverse facilities within walking distance, and therefore the corresponding cluster icons denote areas which have poorer accessibilities. Green denotes



(a) Barcelona four clusters

(b) Barcelona eight clusters

(c) London four clusters

(d) London eight clusters

Figure 2: Examples of K-means clustering in Barcelona and London

higher feature values or more diverse facilities within walking distance, and therefore the corresponding cluster icons denote areas which have richer accessibilities. Diversity thus increases gradually from red to green clusters. Following this clustering step, we can take a centroid in a poor cluster, compare its categories with centroids in richer clusters and make recommendations for category addition to improve its diversity.

4 Future Work

As true for any crowd-sourced dataset, we do not expect the Google Maps data to be exhaustive. But given the extensive coverage of Google Maps in terms of cities worldwide, this is an excellent data source for *scalable* urban analysis. In cities where other data sources are available, like government collected ordnance data or other online map data like Foursquare or OpenStreetMap, these can be used to augment the Google Maps dataset, which we intend to do as part of our future work.

An interesting analysis to be done in future, is informing planning depending on whether a city is mono or polycentric. (Bawa-Cavia 2011) uses Foursquare checkins to identify highly popular urban areas or urban centers in London, New York and Paris. (Batty 2011) uses the subway ticketing data in London to identify urban mobility hotspots. However, Foursquare data is sparse and subway ticketing data is proprietary and difficult to collect for a large number of cities. Owing to the good coverage of Google Maps, our poly-centricity anaylysis can therefore compare multiple cities around the world, potentially enhancing the scalability of prior studies on urban centers.

Finally, our extensive dataset can also help us determine how our cities around the world fare against each other in terms of accessibility indeces. We seek to compare walkability between European and American cities, as explored in prior works (Buehler 2014; Litman 2002), and measure indeces in developing countries to quantify accesibility problems. We envision to replicate a wide variety of independently conducted earlier studies and match their results, while providing insights for the many unexplored cities (those in continents such as Asia, Africa and South America), which have received little or no attention before.

5 Conclusion

Using a scalable methodology, we have gathered web data about urban accessibility and put it to use for answering traditional questions in the urban planning field. We have shown how municipal authorities might profit from crawlable web data to inform evidence-based urban interventions. The private sector might benefit too. For example, since accessibility is associated with quality of city life, websites offering house search (e.g., walkscore.com) might integrate our methodology into their products.

Overall, our proposed methodology for scalable data collection has the potential to study cities around the world, especially those in the developing countries in Asia and Africa, which have been neglected in the literature so far.

References

Batty. 2011. Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PlosONE*.

Bawa-Cavia, A. 2011. Sensing The Urban: Using location-based social network data in urban analysis. In *Pervasive Urban Applications (PURBA)*.

Buehler, R. 2014. 9 Reasons the U.S. Ended Up So Much More Car-Dependent Than Europe.

Cranshaw, J.; Schwartz, R.; Hong, J.; and Sadeh, N. 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Litman, T. 2002. The costs of automobile dependency and the benefits of balanced transportation. *Transportation Research Board* 8(00981871).

Quercia, D.; Aiello, L. M.; Schifanella, R.; and Davies, A. 2015. The digital life of walkable streets. In *Proceedings of ACM International Conference on World Wide Web (WWW)*.

Vaca, C. K.; Quercia, D.; Bonchi, F.; and Fraternali, P. 2015. Taxonomy-based discovery and annotation of functional areas in the city. In *AAAI International Conference on Weblogs and Social Media (ICWSM)*.