Using Semantic Signatures for Social Sensing in Urban Environments

Krzysztof Janowicz

STKO Lab, University of California, Santa Barbara, USA

The Role of Big Data in Transportation Planning, Oct. 2015
Mining Patterns?
Foursquare Check-ins Show ‘The Pulse Of Cities’

‘Every day, millions of people check in on Foursquare. We took a year’s worth of check-ins in Istanbul, London, Chicago, Tokyo and San Francisco and plotted them on a map. Each dot represents a single check-in, while the straight lines link sequential check-ins. What you can see here represents the power of check-in data — on Foursquare, every city around the world pulses with activity every hour of every day.’

https://foursquare.com/infographics/pulse
Foursquare Check-ins Show ‘The Pulse Of Cities’

- (I) The videos are **not interactive**, e.g., one cannot click on any of the check-in events or places to gain additional insights.
- (II) The videos are rendered based on a **fixed geographic scale** and focused on a particular part of the city. Thus, one cannot pan or zoom.
- (III) The millions of check-ins are aggregated to a **single** non-specific **day**, thus hiding well known patterns, e.g., weekdays versus weekends.
- (IV) Foursquare’s POI taxonomy consists of >400 POI types grouped into 9 top classes. While such a generalization is necessary, it is unclear how and why certain types are **categorized** in specific ways (e.g., *Cemeteries* as *Great Outdoors*). A **crisp class** membership on such a coarse level will necessarily introduce arbitrary decisions and thus will significantly alter the observed temporal pulse of the city.
- (V) Similar to other UGC, Foursquare contains data of widely **varying quality**. E.g., users often type their own houses as *Castle* or check-in to features of the types *Road*, *Trail*, or *Taxi*. While this is a consequence of UGC, it is important to clean the data.
Semantic Signatures
Whether on land or in space, observatories and their sensors serve different purposes and are most useful when they work together.
Spectral Signatures, Bands, and Remote Sensing

- **Spectral signatures** are the combination of emitted, reflected, or absorbed electromagnetic radiation at varying wavelengths (bands) that uniquely identify a feature type.

- **Spectral libraries**, the idea of sharing spectral signatures, has revolutionized remote sensing.
Semantic Signatures As Analogy To Spectral Signatures

- **Geospatial bands**
  - based on geographic location
    - ANND
    - Ripley’s K Bins
    - J Measure
    - $D_{zero}$
- **Temporal bands**
  - based on geo-social check-ins
    - 24 Hours
    - 7 Days
    - Seasons
- **Thematic bands**
  - based on venue tips and reviews
    - LDA topics
    - TF-IDF
- Makes use of data heterogeneity
Semantic signatures and bands are an analogy to spectral signatures.

So far, we have mined and modeled hundreds of bands for hundreds of different geographic features on the micro, meso, and macro-scale.

Successfully applied signatures to categorization, deduplication, semantic enrichment, cleansing, visualization, exploration, reverse geocoding, ontology alignment,...
Thematic Bands & Geo-Indicativeness

Santa Barbara, California

From Wikipedia, the free encyclopedia

This article is about the city. For the island, see Santa Barbara Island. For the county, see Santa Barbara County, California.

Santa Barbara is the county seat of Santa Barbara County, California. Situated on a south-facing section of coastline, the longest such section on the West Coast of the United States, the city lies between the steeply rising Santa Ynez Mountains and the Pacific Ocean. Santa Barbara’s climate is often described as Mediterranean, and the city has been promoted as the "American Riviera". As of the 2010 census, the city had a population of 88,410, a loss of 1,190 from the previous census, making it the second most populous city in the county after Santa Maria. While the contiguous urban area, which includes the cities of Goleta and Carpinteria, along with the unincorporated regions of Isla Vista, Montecito, Mission Canyon, Hope Ranch, Summerland, and others, has an approximate population of 220,000. The population of the entire county in 2010 was 423,895.

In addition to being a popular tourist and resort destination, the city economy includes a large service sector, education, technology, health care, finance, agriculture, manufacturing, and local government. In 2004, the service sector accounted for fully 35% of local employment. Education in particular is well represented, with five institutions of higher learning on the south coast (the University of California, Santa Barbara, Santa Barbara City College, Westmont College, Antioch University, and the Brooks Institute of Photography). The Santa Barbara Airport serves the city, as does Amtrak. U.S. Highway 101 connects the Santa Barbara area with Los Angeles to the southeast and San Francisco to the northwest. Behind the city, in and beyond the Santa Ynez Mountains, is the Los Padres National Forest, which contains several remote wilderness areas. Channel Islands National Park and Channel Islands National Marine Sanctuary are located approximately 20 miles (32 km) offshore.

Places at geographic location 34.43, -119.71 are:

- of types city, county seat, ...
- at the coastline, near the mountains, have Mediterranean climate, ...
- described in terms of urban area, economy, tourism, government, employment, ...

**Interesting observation**: some of these terms will co-occur by type, others per region.
Thematic Bands & Geo-Indicativeness

- A thematic band can be computed out of unstructured text from sources such as Wikipedia, travel blogs, news articles, and so forth.

- Non-georeferenced plain text is often still geo-indicative

- Different types of geographic features have different, diagnostic topics associated to them (out of 500 topics)

- Indicative topics and be lifted to the type-level.

- Here, we modeled topics using latent Dirichlet allocation (LDA)
**Thematic Bands & Geographic Feature Types**

- **City** topics: 204 > 450 > 104 > 282 > 267 > 497 > 443 > 484 > 277 > 97 > ...
- **Town** topics: 425 > 450 > 419 > 367 > 104 > 429 > 266 > 69 > 204 > 308 > ...
- **Mountain** topics: 27 > 110 > 5 > 172 > 208 > 459 > 232 > 398 > 453 > 183 > ...
Study **geo-social** check-in data to location-based social networks.

- Aggregate them to the **feature type** level and clean them.
- Intuitively, people visit **wineries** in the afternoon and evening and **bakeries** in the mornings.
- Combining weekly and hourly **bands** to create place type **signatures**.
**Spatial Bands**

- POI plotted by *similarity* to bar and post office in OpenStreetMap data (London)
- **Similarity** measured as association strength in OSM change history
- **Bars** (and similar features) tend to *clump* together
- **Post Offices** (and similar features) are rather *uniformly* distributed
**Spatial Bands**

- $D_{zero}$ measures the **likelihood** of features of a certain **type** to co-occur within a specific **semantic and spatial range**.

- General idea: generate **recommendations** and **clean up** data based on **type likelihood**. 'How likely is a post office directly next to an existing one?'
(Remote sensing) **sensors** can be characterized by their **resolution**

- **Spatial resolution**: smallest feature that can be detected, i.e., the pixel size.
- **Temporal resolution**: smallest time interval between a repeated observation.
- **Spectral resolution**: number, position, and width of spectral bands.
- **Radiometric resolution**: small distinguishable differences in radiation magnitude.

Analogous **social sensor resolutions**, e.g., types of bands, number of topics.
Circular temporal signatures histograms for **Theme Park** (a,b,c) and **Drugstore** (d,e,f). About 50% of ≈ 400 Point Of Interest (POI) types are **regionally invariant** in the **USA**.
The ‘Foursquare-day’

How and when do people check-in at places, manually, automatically?

Do they check-out? If not, after what time are they checked-out automatically?
Which place types can be **meaningfully** distinguished (in DBpedia)?
The pulse of

Los Angeles
POI Pulse: Explore the Pulse of Los Angeles

http://poipulse.com/
The POI Pulse Implementation (All Venues)
The POI Pulse Implementation (Upper-level Categories)
The POI Pulse Implementation (Switching to Vectors)
The POI Pulse Implementation (User Interaction)
The POI Pulse Implementation (Burst Mode)
POI Pulse uses semantic signatures to model **default behavior**. In combination with the **burst mode**, this can be used for a variety of tasks such as detecting outliers, emerging event, or changes in places.

The **memory footprint** of the semantic signatures is minimal, yet they can be used for highly complex tasks.

So far, we **successfully applied** signatures to categorization, deduplication, semantic enrichment, cleansing, visualization, exploration, reverse geocoding, ontology alignment, and other tasks.

The semantic signatures analogy provides us with a well-known and tested framework (sensors & observations) to serve as a formal underpinning for **social sensing**.

We hope that a **semantic signature library** will be equally transformative to social sensing as spectral signature libraries have been to the natural sciences.