POI Pulse: A Multi-Granular, Semantic Signatures-Based Information Observatory for the Interactive Visualization of Big Geosocial Data

Grant McKenzie[1], Krzysztof Janowicz[1], Song Gao[1], Jiue-An Yang[1,2] and Yingjie Hu[1]

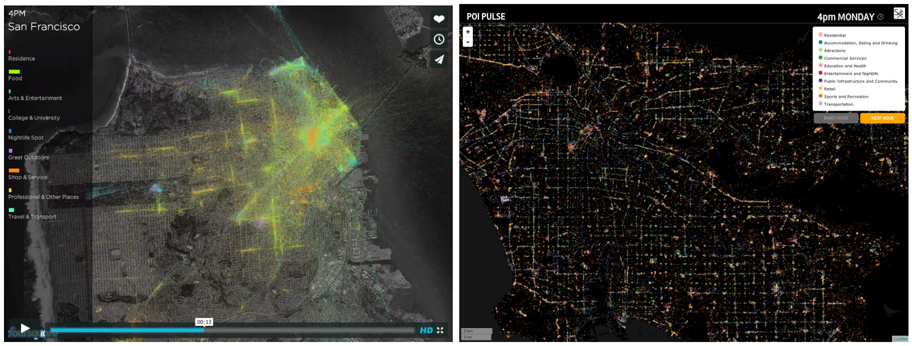
[1]STKO Lab, Department of Geography, University of California, Santa Barbara, USA
[2]Department of Geography, San Diego State University, USA

Abstract

The volume, velocity, and variety at which data are now becoming available allow us to study urban environments based on human behavior at a spatial, temporal, and thematic granularity that was not achievable until now. Such data-driven approaches opens up additional, complementary perspectives on how urban systems function, especially if they are based on User-Generated Content (UGC). While the data sources, e.g., social media, introduce specific biases, they also open up new possibilities for scientists and the broader public. For instance, they provide answers to questions that previously could only be addressed by complex simulations or extensive human participant surveys. Unfortunately, many of the required datasets are locked in data silos that are only accessible via restricted APIs. Even if these data could be fully accessed, their naïve processing and visualization would surpass the abilities of modern computer architectures. Finally, the established place schemata used to study urban spaces differ substantially from UGC-based Point of Interest (POI) schemata. In this work, we present a multi-granular, data-driven, and theory-informed approach that addressed the key issues outlined above by introducing the *theoretical and technical* framework to interactively explore the pulse of a city based on social media.

1 Introduction and Motivation

Today's data universe offers access to a plethora of data at a spatial, temporal, and thematic resolution unthinkable just a few years ago. This data revolution is accompanied by the emerging 4th paradigm of science [8, 6] in which synthesis is the new analysis. Those changed realities cast off visions of *information observatories* [18]



(a) Screenshot from *Foursquare* video (b) *POI Pulse* interactive visualization

Figure 1: The pre-generated video (a) and the interactive *POI Pulse* system (b).

in which complex systems, such as urban spaces[1], could be observed and better understood based on exploiting the variety, volume, and velocity of Big Data [12, 11]. Those, however, who tried to explore these new possibilities often encountered equally big challenges. First, major parts of Big Data still reside in closed proprietary silos with limited API access. Second, the metadata, e.g., provenance, and conceptual schemata required for any serious use by scholars are often not present, intransparent, or differ substantially to those established in science. Finally, the sheer volume and velocity makes interacting with or even just visualizing the data difficult to say the least.

For many of us, an information observatory for urban spaces in which user-generated *real-time* content reveals spatial, temporal, and thematic patterns and traits of human behavior, is a tempting idea as it aligns well with the Digital Earth vision. Consequently, a posting[2] on Foursquare's infographics blog in October 2013 raised a lot of attention. It linked to a series of videos showing the *pulse* of different cities such as San Francisco. The animations were entirely derived from mining massive amounts of user check-ins to the Foursquare Location-based Social Network and were aggregated to a single virtual day; see Figure 1a.

While the visualization itself is quite stunning, the Foursquare videos have several shortcomings: **(I)** The videos are not interactive, e.g., one cannot click at any of check-in events or places to gain additional insights.[3] **(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 more than 400 POI types grouped into 9 top-level classes (see Figure 1a). While such generalized classes are necessary and useful, it is not clear how they were derived nor why certain POI types are categorized in specific ways. Furthermore, a binary class membership on such a coarse level will

[1] See `http://www.urbanobservatory.org/compare/index.html` for an early example.

[2] https://foursquare.com/infographics/pulse

[3] Interested readers may try to find an explanation for the moving fast *Food* cluster in San Francisco at 4am; see `https://foursquare.com/infographics/pulse#san-francisco`.

necessarily introduce arbitrary decisions and thus will significantly alter the observed temporal pulse of the city. For instance, *Cemeteries* are categorized under the *Great Outdoors* category. **(V)** Similar to other UGC, Foursquare contains data of widely varying quality. For instance, users often classify their own houses as *Castle* or check-in to features of types *Road*, *Trail*, or *Taxi*. While this is a consequence of UGC, it is important that the data be cleaned.

Inspired by Foursquare's pulse videos and the theoretical and technical limitations of interacting and visualizing Big Data, we decided to address the aforementioned restrictions by designing a *POI Pulse* portal for Los Angeles;[4] see Figure 1b. Naturally, as scientists we are more interested in those theoretical and technical aspects than the application as such, but we will use it as the joint **leitmotiv** that connects the following **research questions** which make up the scientific contribution of this work:

$\mathcal{R}1$: Given the $>$400 POI type defined by Foursquare users, is it possible to derive an alternative top-level classification that is informed by existing and well established POI schemata (e.g., defined by Ordnance Survey) and still true to the original Foursquare data and user-behavior?

$\mathcal{R}2$: Most likely, the reason for showing a pre-rendered video is the fact that even the most modern Web bowser using HTML5, CSS3, and effective JavaScript engines, cannot render the hundreds of thousands of POI as vectors thus making interaction cumbersome. Is it possible to use a scale-dependent, seamless combination of raster and vector tiles to render approximately 200,000 POI for Los Angeles, and still make the interface interactive and responsive? What is the tipping point from which vector tiles will be faster than raster tiles?

$\mathcal{R}3$: Given the legal API limits of closed data silos such as Foursquare, can we generalize check-ins, individual POI, and their attributes, e.g., tips, to a type-level *default behavior* that allows us to model the pulse of a city with minimal data requirement? Is it possible to seamlessly switch to a real-time, *burst* mode at zoom scales that do not exceed the daily API limits and thus also give access to real time data?

$\mathcal{R}4$: Can we improve on the Foursquare baseline by offering a pulse for all hours of the full week instead of a single day? Can we show binary upper-level categories but seamlessly switch to a more nuanced view at a reduced zoom level to show a probabilistic category membership?

In the following, we present a multi-granular, data-driven, and theory-informed[5] approach that addresses these research questions by introducing the *theoretical and technical* framework to interactively explore the pulse of a city based on social media.

2 A Data-Driven and Theory-Informed POI Taxonomy

In this section, we discuss how to derive a POI taxonomy by combining data-driven techniques with existing top-down classification schema. Many different POI vocabularies, taxonomies, and schemata have been defined in the past few years, e.g., schema.org, the Ordnance Survey POI classification system, the OpenStreetMap map features, OpenCyc, the Linked Geo Data ontology, the GeoNames ontology, or even

[4]Explore the portal at `http://www.poipulse.com`

[5]I.e., including existing top-down schemata from the research literature and industry.

WordNet, to name a few. Unfortunately, most of these are not suitable for our purpose. Sources such as WordNet are not specific enough, while platforms, such as OpenCyc, introduce distinctions (e.g., *man-made structure)* that are interesting from an ontological perspective but hinder the task at hand. OpenStreetMap is notorious for its flat key-value pair classification and also introduces many feature types that are not POI specific. Similarly, the GeoNames feature classes are not suitable, since all POI types defined in Foursquare would end up in the same class (*S spot, building, farm*). Consequently, after an initial review, we decided to use schema.org.

Schema.org is a data markup ontology jointly constructed by Google, Yahoo, Microsoft, Yandex, and the W3C. Intuitively, one would assume that such an ontology is most suitable to provide an upper-level abstraction for the >400 Foursquare types and should be able to replace the 9 top-level classes. One may also expect that schema.org was developed with datasets such as Foursquare, Yelp, etc, in mind. Surprisingly, however, that turned out not to be the case. For instance, schema.org distinguishes between *Places* and *Organization* as one of its top-level distinctions. While this is not wrong, the fact that Internet cafes are considered organizations but movie theaters are places is surprising.[6] Due to many similar cases and ontological decisions taken by schema.org, it became clear that we needed another classification.

Eventually, we selected the Ordnance Survey (OS) POI classification system (v. 2.3) [16]. In contrast to Web and data-driven resources, the OS classification is an administrative and UK-specific resource. The OS system consists of 9 classes at the 1st level, 49 classes at the 2nd level, and 600 POI types at the 3rd level. We are only interested in the first level here (OS1). It consists of the following classes: ***01*** *Accommodation, eating and drinking*, 0***2*** *Commercial Services*, ***03*** *Attractions*, ***04*** *Sport and entertainment*, ***05*** *Education and health*, ***06*** *Public infrastructure*, ***07*** *Manufacturing and production*, ***09*** *Retail*, and ***10*** *Transport*. We will use this classification as our top-down, theory-informed POI schema and in the following section describe how to use data-driven techniques to semi-automatically align the Foursquare types to this schema.[7]

2.1 Multi-dimensional Characterization of POI Types

The variety of big data presents new possibilities to understand POI from different perspectives. In previous work, we proposed the concept of *semantic signatures* to characterize a place using spatial, temporal, and thematic patterns [9]. As an analogy to *spectral signatures* in remote sensing, *semantic signatures* differentiate types of places based on multiple *bands*. In this work, we employ the *semantic signatures* idea, and extract a number of descriptive dimensions from the Foursquare data to characterize POI.

[6] Via: Thing >Organization >LocalBusiness >InternetCafe (see `http://schema.org/InternetCafe`) and Thing >Place >CivicStructure >MovieTheater (see `http://schema.org/MovieTheater`).

[7] To improve readability, we will refer to the Foursquare classes as *POI types* and to the OS1 classes as *upper-level classes*.

2.1.1 Temporal Bands

The temporal bands are derived from 3,640,893 check-ins to 938,031 venues from 421 Foursquare categories in Los Angeles, New York City, Chicago, and New Orleans. These check-ins have been collected for 4 months starting October 1st, 2013. Consequently, we cannot use them to understand seasonal effects but focus on the 168 hours of the week. The temporal resolution of the data is 2 hours, i.e., while we have hourly check-in times, the duration of check-ins is unknown and users are automatically checked out after 2 hours. In our work, we are neither interested in the particular venues, check-ins, nor users,[8] but in studying the temporal **default behavior** of users towards **types** of POI. In other words, we are interested in the fact that bars are visited in the evenings and especially during weekends, while universities are mostly visited during the workdays between 7am-5pm. Figure 2, depicts 168 bands that jointly form the temporal signature for three POI types. The data represents probability values for check-ins to the given type (by hour bins), i.e., the 168 bands sum up to 1. Despite the large sample, we had to remove outliers as some of the POI types, e.g., *Molecular Gastronomy Restaurant*, have fewer venues than others. We used 4 standard deviations from the mean as cutoff. While we have not used these temporal bands before, we applied a coarser and more limited temporal signature to predict types for untagged POI successful [19]. Thus, we expect the temporal bands to play a major role in the derivation of the POI taxonomy.

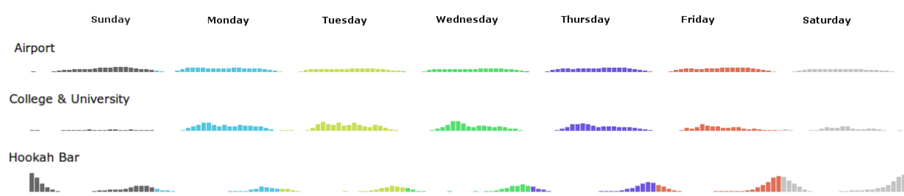


Figure 2: The weekly temporal bands for selected POI types by hour.

2.1.2 Thematic Bands

A representative subset of the venues (274,404) from the 421 Foursquare categories (POI types) have been used to derive another, yet very different set of bands that will jointly form the thematic signature; cf. [17]. We collected all user-contributed tips for those venues, stemmed all words, generated venue specific documents out of them, grouped these documents by POI type, and then used Latent Dirichlet Allocation (LDA) [1][9] to extract topics. LDA is an unsupervised, generative probabilistic model used to infer latent topics in a textual corpus. We trained LDA by treating the tips associated with all venues of a given type as single documents. LDA uses a *bag-of-words* approach to uncover topics that are represented as multinomial distributions over words. Each topic is composed of multiple words and their relative importance

[8]Even more, due to API restrictions these data should not be stored for more than 24 hours.

[9]Due to the API restrictions, we are only storing the derived latent topics per POI type.

for this topic. Figure 3 uses word clouds to visualize the top 18 words in three topics by scaling them according to their probability. It is important to note that each stemmed word extracted from the tips appears in each topic with a different probability. LDA topics do not necessarily correspond to themes typically formed by humans.

(a) Topic 39 (b) Topic 26 (c) Topic 53

Figure 3: Words that make up LDA topics scaled by their relative probability.

Topic 53, for example, is interesting as it prominently contains Spanish terms, while topic 26 is related to terms about markets, flowers, plants, etc. We are **not** interested in the specific terms but only their indicativeness, i.e., how diagnostic they are in predicting the type of place. For instance, topic 53 is more **likely** to appear in relation to POI *types* such as *Mexican Restaurant* than within tips contributed to *Yoga Studios*.

2.1.3 Spatial Bands

The spatial distribution patterns of POI types in urban areas differ. To achieve a more holistic signature of the POI, we also introduce 14 spatial bands. The first set of bands is derived from the average of nearest-neighbor distances (ANND) among all POI typeS. The values have been normalized to [0-1] such that the larger value indicates dispersion while the smaller value represents clustering. The next set of bands are derived from Ripley's K which offers the potential for detecting both different types and scales of spatial patterns. The K measure computes the average number of neighboring venues (of the same type) associated with each POI within a given distance and then compares them to the expected value under completely spatial randomness. We chose 10 distance thresholds and calculated the corresponding Ripley's K measures as 10 spatial bands for all POI types. Figure 4 shows that the K measure helps to evaluate how the spatial clustering or dispersion pattern of each POI type changes when the neighborhood distance changes. For instance, The values of ANND *Police Station* (0.721) and *Night Club* (0.702) are very close, while their spatial clustering patterns are different at multi-distance bands (scales). Both ANND and Ripley's K measures only consider the distance or the number of neighboring venues but ignore the POI type information for spatial point pattern analysis. In urban areas many POI types (such as nightclubs and bars) often clustered together. The different types of spatial mixture patterns should also be taken in to consideration. To address this issue, we introduce a third family of bands called the *J Measure*. The *J Measure* involves generating triangles between all POI of the same type, counting the number of other distinct POI types within each triangle and dividing it by the total number of POI types. We computed the mean, median, and standard deviations for the *J Measure* for all POI types. For instance, the

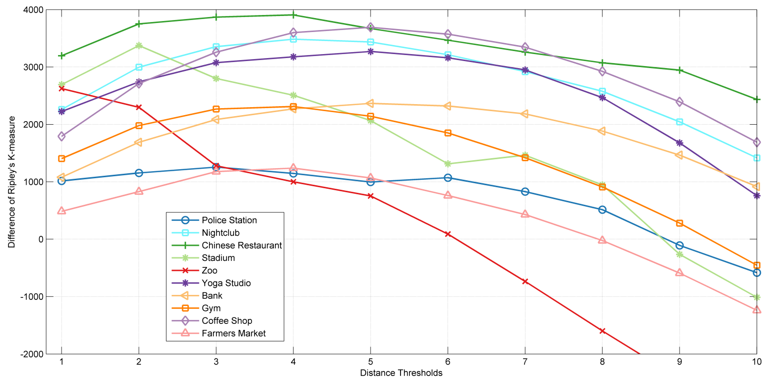


Figure 4: Ripley's K for 10 types (The y-axis value represents the difference between the observed value of K-measure at a given distance and the expected value under the CSR simulation process.)

mean *J Measure* for *Police Station* (0.257) is larger than that of *Night Club* (0.176), which indicates a larger POI type diversity between adjacent police stations.

2.2 Data Cleaning

As a next step we cleaned the dataset by removing all POI types that either refer to clearly linear features or were overly generic. Examples include types such as *Road* and *Trail*, and non-descriptive types such as *Building* or *City*. We also removed types that are a pure artifact of UGC and that we know have no instances in the Greater Los Angeles, e.g., *Volcano*. Similarly, we removed the type *Castle*, assuming that the 77 POI within the dataset are from user's that took the "my home is my castle" motto too literally. Finally, we removed clearly non-stationary POI such as *Plane* and *Taxi*, while leaving the *Food Truck* type in the dataset. This reduced the number of Foursquare POI types from 421 to 387 and the LA POI dataset from 178,814 POI to 164,902.[10]

2.3 Information Gain

Given the 246 different bands that jointly form our semantic signatures, it is interesting to discover which of these bands are most diagnostic in terms of their ability to estimate the membership of a particular POI type with respect to an upper-level class. This is for two reasons: First, it allows us to reduce the high-dimensional space by excluding dimensions/features that do not contribute (much) to the classification; Second, it provides intuition about the expected bias, and, thus, the limits of data-driven classification. For instance, if most of the thematic bands were not diagnostic, it would be difficult to tell apart the *Airport* type from the *Emergency Room* type as the two share similar temporal signature, i.e., people visit them during all hours and days of

[10]Our original dataset contained 191,998 POI but this included uncategorized POI as well.

the week. Hence, both types would more likely be classified as belonging to the same class, e.g., *Education and Health*, while they should belong to two distinct classes namely *Transport* and *Education and Health*.

Information gain is a measure of the expected decrease in entropy [15]. It provides an assessment of the contribution of a particular feature (i.e., a specific band) for predicting the dependent variable, i.e., the upper-level class. To compute the information gain of the 246 bands, we jointly agreed on a set of POI types considered as clear matches for the respective OS1 classes. For instance, *German Restaurant* was manually classified as being a subtype of *Accommodation, Eating and Drinking*. Next, the information gain for all (discretized) bands was computed using this training set and the median and arithmetic mean scores were determined. Assuming that simpler models can better capture the underlying structures [15], all bands with information gains scores below the mean were removed, leaving 159 bands that were considered diagnostic.

Band	Information Gain	Band	Information Gain
temp143	0.772	temp161	0.695
temp59	0.750	temp88	0.693
temp107	0.744	theme39	0.519
temp60	0.725	spatial4	0.234
temp35	0.712	temp29	0.034

Table 1: The 7 overall most diagnostic bands according to their information gain, the most diagnostic thematic and spatial bands, and the least diagnostic band.

Table 1 shows some results. It is interesting to note that all top bands are temporal. In fact, the first non-temporal band ($theme39$) is ranked 56th. This thematic band is graphically represented in Figure 3a. The first spatial bands ($spatial4$) is ranked 134th. Examining the top temporal bands shows that the typical lunchtime hours (11am-12pm), close of business hours (4-5pm), and dinner/nightlife hours (10-11pm) are most relevant, as is the distinction between workdays and weekends. Band $temp143$, for instance, corresponds to Friday 11pm while the least diagnostic band ($temp29$) corresponds to Monday 5am. Consequently, while all 159 bands will contribute to the classification, we can expect the classifier to have more difficulties in learning the membership for classes such as *Public Infrastructure* that consist of POI types with widely varying temporal bands, e.g., *Police Station* versus *City Hall*. This will result in lower precision and recall values for such upper-level classes and will be discussed in the following sections. One could, of course, consider and extract additional bands. However, this is out of scope for the paper at hand and significantly restricted by the availability of attribute data from typical POI data sources.

2.4 Interactive Classification

The creation of bands and their reduction via information gain sets the stage for classifying the POI types from Foursquare using the Ordnance Survey level 1 classes. To do so we used a combination of machine learning and manual corrections in two different runs.

First, we selected the previously generated training set of POI types and trained a Support Vector Machine (SVM) [4] with a polynomial kernel. Next, we **predicted** the OS1 classes of all POI types using the same training set. We check all cases where the assigned and the predicted classes varied and decided manually which class to use. Interesting examples where we changed our initial decision include *Bagel Shop* that we initially classified as *Retail* which is rather a breakfast place (thus, *Accommodation, eating and drinking*) in the US. Similar cases included *Brewery*, *Nail Salon*, and other POI type that could be categorized as belonging to different classes. Another good example is all college buildings. For instance, should *College Football Field* be categorized as *Education and Health*, *Sports and Entertainment*, or *Attractions*? From the point of view of a social check-in application such as Foursquare, the number of users that view a football field as an attraction is orders of magnitude above the actual players (for which the football fields should belong to the sports class). We will address this multi-class nature of many POI types from a visual perspective in section 4.

Finally, we trained the SVM with the new training set and subsequently with all POI types. We computed the recall and precision for this run and manually inspected all mismatching class predictions. This led to some interesting findings about the bias in the Foursquare data, its crowd-sourcing nature in contrast to the administrative OS level 1 classes, as well as socio-political differences between the US-based type data and the UK-based schema. For instance, according to the OS classification *Recycling Facility* should be categorized as *Public Infrastructure* while they are *Commercial Services* in the US. Other interesting cases included *Public Art* that SVM successfully categorized as *Attraction*, or *Tailor Shop* that was predicted to belong to *Retail* (but could also have been a *Commercial Service*).

Target Class	F1	Precision	Recall
Accommodation, eating and drinking	0.8343	0.869	0.8022
Attractions	0.6479	0.561	0.7667
Commercial Services	0.5882	0.6667	0.5263
Education and health	0.7792	0.7692	0.7895
Entertainment and Nightlife	0.8235	1	0.7
Public Infrastructure and Community	0.5946	0.6471	0.55
Retail	0.8611	0.8267	0.8986
Sports and Recreation	0.7568	0.6774	0.8571
Transport	0.6957	0.7273	0.6667

Table 2: F-score, Precision, and Recall for upper-level classes after the 2nd run.

After inspecting the predicted class membership probabilities for each single type, we realized that based on the nature of the Foursquare POI types as well as the previously mentioned bias in our 159 most diagnostic bands, we would need to change some of the OS level 1 classes. We decided to remove the *Manufacturing and production* class as it only has three subtypes in our dataset, renamed *Public infrastructure* to

Public Infrastructure and Community to also include religious places, and finally split *Sport and entertainment* into two distinct classes: *Sport and Recreation* as well as *Entertainment and Nightlife*. As the temporal bands were found to be the most diagnostic features in our dataset and as we want to show the pulse of a city by hours and days, a joint class for POI types such as *Basketball stadium*, *Martial Arts Dojo*, and *Strip Club* was not feasible.

Target Class	A.E.D.	Attr.	Comm.	Edu.	Entert.	Public	Retail	Sports	Trans.	#
Accommodation, Eat, Drink	73	6	0	1	0	0	6	5	0	91
Attractions	0	23	0	0	0	0	4	3	0	30
Comm. Services	0	4	20	5	0	3	1	2	3	38
Education, Health	0	0	3	30	0	3	0	2	0	38
Entertainment, Nightlife	10	0	0	1	28	0	0	1	0	40
Public Infrastructure, Community	0	2	1	1	0	11	1	4	0	20
Retail	0	3	1	0	0	0	62	3	0	69
Sports, Recreation	1	1	3	1	0	0	1	42	0	49
Transport	0	2	2	0	0	0	0	0	8	12
#	84	41	30	39	28	17	75	62	11	387

Table 3: Confusion Matrix after final class predictions.

The second run consisted of a new training set based on the new upper-level classes and the lessons learned from the first run. We trained a SVM and predicted class membership for the training set as well as all other POI. The F-score, precision, and recall for this 2nd run are listed in Table 2. While the results for the new *Entertainment and Nightlife* class or the OS1 class *Accommodation, eating and drinking* are *very high*, other classes are more difficult to predict. This is largely due to the heterogeneity within such classes as well as the fact that some POI types cannot be distinguished based on the temporal, thematic, and spatial signatures. The class *Public Infrastructure and Community* offers good examples of this, and thus, has a relatively low F-score. The class includes POI types such as *Police Station*, *City Hall*, and *Mosque*, that vary substantially with respect to all bands. Finally, some POI types would require very different bands for their successful classification, e.g., sentiment analysis could be used to better distinguish police from fire stations. Table 3 shows a confusion matrix to give an overview of the varying classification success.

Figure 5 shows a fragment of a Multi-Dimensional Scaling plot. Each node corresponds to a types, while colors indicate class membership. The lines represent the top 20 % most similar pairs, while the node sizes indicate Kruskal stress. Classes such as *Accommodation, eating and drinking* (blue) and *Entertainment and Nightlife* (yellow) form densely connected clusters while other classes, e.g., *Public Infrastructure and Community* (pink) are less coherent. This essentially confirms our findings visually.

Summing up, we derived a new upper-level classification schema based on an existing top-down schema as well as a data-driven way in which we let the data speak for itself to inform (and in most cases decide on) our final classification.

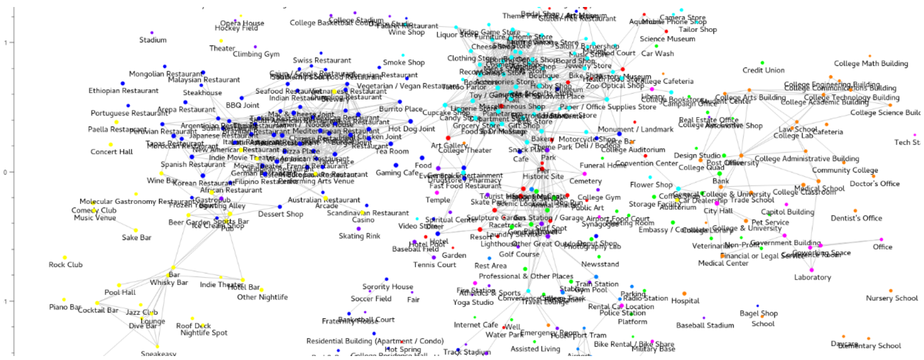


Figure 5: Fragment of a Multi-Dimensional Scaling plot showing the uper-level classes as colors; Kruskal stress: 0.258.

3 Interaction and Visualization – Rasters vs. Vectors

Once the upper-level classes were established and the POI types were successfully classified, the focus shifted to the challenge of visually rendering the over 170,000 individual POI. One of the primary issues that continues to plague web mapping and cartography, is the speed at which data can be displayed. Recent advances in browser technology have allowed for dramatic changes in the way data can be processed and visualized. From a web mapping perspective, this increased reliance on browsers allows for improved interaction with data, reducing the need for continual client-server requests. Given the large amount of data and interactivity needs of this project, optimization of the web mapping component is essential. This section presents an efficient method for visualizing this big geosocial data.

Since a true city pulse requires equal contribution from all POI, an early decision was made not to cluster or reduce the POI when constructing the visualization. This created a challenge in determining an efficient means for visualizing approximately 170,000 points through a web browser. The state-of-the-art for many years has been to serve a collection of static image tiles pre-rendered by a mapping toolkit. The structure of these tiles typically follows a simple coordinate system. Each tiles has a Z (map scale) coordinate and an X and Y coordinate that describe its position within a square grid. For every Z-level increase, the number of tiles required increases by a factor of four, leading to extremely large tilecaches, depending on the number of zoom levels and extent of the area of interest. For reasons of practicality, most mapping applications restrict the number of tiled zoom levels to 20. Zoom level 0 represents the entire world in a single tile while level 19 projects the Earth at a map scale of 1:1,000. The power of the image tiling scheme is that the size of the image file transmitted to the client is minimally influenced by the size of the data.

Recent W3C standards, such as *HTML5*[11] and *Scalable Vector Graphics (SVG)*[12], combined with powerful modern web browsers continue to push the boundaries of

[11] http://www.w3.org/TR/html5/
[12] http://www.w3.org/TR/SVG11/

what can be done in web cartography. While image tiles allow cartographers to solely *display* content via the web, *Vector Tiles* offer users the enhanced ability to interact directly with the content. Vector tiles take a similar approach as image tiles in that they divide data in to smaller sizes in order to enable faster loading times leveraging modern browser parallelization and asynchronous data requests. Unfortunately, the enhancement of offering direct data interaction also increases the burden on the client side as data rendering is now being executed locally. Thus, the goal is to find the *tipping point* at which a web mapping framework should switch between raster and vector tiles.

3.1 The Tipping Point

Ideally, vector tile representations of POI should be displayed at all map scales allowing for maximum interaction with the data. An experiment was run on three different networks in which both Vector and Raster representations of the POI dataset were loaded. Each tile format was loaded 200 times at each of the zoom levels between 10 and 16. The loading times (in milliseconds) were recorded and averaged and the results are shown in Figure 6. As one can see, the loading time required to display all POI (Zoom level 10) in vector format is simply not practical. With each increase in zoom level, the transfer/rendering time for the vector tiles decreases. Only those tiles that intersect with the view-port are transferred to the browser and rendered. This means that fewer and fewer points are displayed, reducing the amount of data to be transmit and rendered.

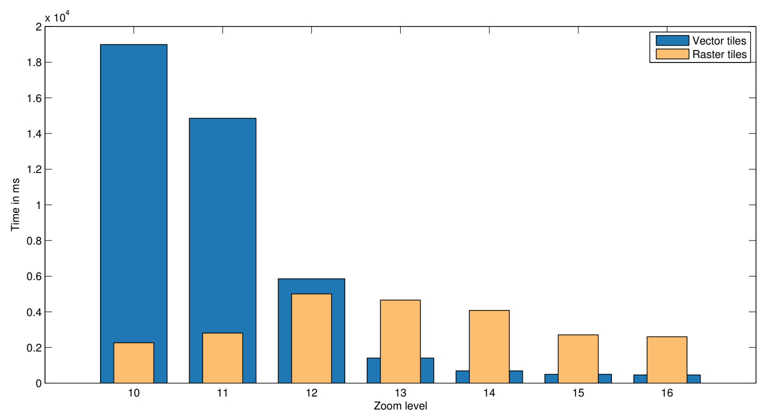


Figure 6: Raster vs. Vector tile loading times (in ms).

Comparatively, the loading times for image tiles (PNG-8) increase between zoom levels 10 and 12 and steadily decrease thereafter. This initial increase can be explained by the number of rendered tiles. If a given view-port requires 32 tiles to cover the entire map display, for example, only 10 of these tiles will contain data at zoom level 10. This means that 22 tiles need not be transferred from the server or rendered by the client. As zoom levels approach 12, the number of tiles that render content increases until all 32 visible tiles contain some amount of POI. The reduction in loading times between zoom levels 12 and 16 can then be explained by the amount of data rendered in each PNG. As the map scale increases, the total number of POI visible in the view-port decreases, indicating that the average number of POI rendered in each PNG approaches zero.

Given the significant disparity in loading times, the POI can clearly not be rendered solely in vector format. With the purpose of reducing loading time and maximizing user-data interaction, the switch in tile formats should occur between zoom levels 12 and 13. While loading time is a key contributor to this decision, the total number of POI is also relevant. Interacting with POI data at a map scale smaller than 1:70,000 is simply not possible as the ability to select an individual POI at this scale is arduous.

3.2 Technology

Adopting this idea of switching between vector and raster representation, the POI Pulse application is implemented. This platform is built using a novel combination of web technologies. Based in HTML5 and Javascript, the *Leaflet* v0.7 Javascript framework is employed for map display and interaction. The *Data-Driven Documents (D3)*[3] v3.4 library is used for manipulating and rendering data in Scalable Vector Graphics format through Javascript. On the back-end, *TileStache* v1.49 extracts the data from a PostGreSQL 9.2/PostGIS 2.0 spatially enable database, and organizes the POI in the file structure required by Leaflet. TileMill and CartoCSS are employed for cartographic styling exported as XML. *MapNik* v2.2 reads these style documents and renders image tiles while *TopoJSON* [2] v1.4.2 vector tiles are generated through TileStache and rendered on-the-fly with D3 and cascading style sheets (CSS). On page load, the map shows all POI in white with an opacity value used to indicate popularity for each hour of the week. For zoom levels 11 and 12, diverging colors, selected through the Color-Brewer application [7], were assigned to the ten upper-level classes. In order to allow visibility control for each class, ten separate image tile caches were created.

Respectively, vector tiles are not pre-rendered and thus do not require numerous tile caches. Since vector tiles contain the raw geographic location and attributes, these data can be rendered on the client. Toggling the visibility of classes within the vector tiles is handled by iterating through the vectors and changing the visibility parameter for the appropriate SVG element. Restricting the zoom levels at which these tiles are rendered means that a limited number of POI require real-time rendering, but this feature requires the generation of a large number of vector tiles. Remember that the number of tiles required at each zoom level increases by a factor of four as the zoom level increases. Excluding empty tiles, this means that level 13 requires 704 vector tiles while level 14 requires 2,666.

3.3 Pre-loading Map Tiles

Once the image and vector tiles have been generated, implementing them in the user experience becomes the next challenge. The temporal nature of these data require that a new set of tiles be displayed to the viewer with each click of the hour-advancement button; see Figure 1b. Regardless of number of tiles or tile size, the process of adding tiles to the map always takes the web mapping framework a split second to organize the tiles and display them. The most common technique in viewing time series data is to procedurally remove one set of tiles from the map and add another set. In theory, this makes sense, but in practice, this process produces a moment where neither the previous tiles nor the new tiles are visible on the map. This creates what psychologists refer to as a *mask* between experiment tasks, removing any link between the previous image and the next. Unfortunately this has a negative impact on the application's user experience. Since the changes in activity are quite small from one hour to the next, this masking effect overpowers the visual effect of the minute changes necessary to understand the urban pulse. In order to circumvent this issue, we pre-render tiles on the client and set the opacity value of zero. Initially an array of four hours of data are loaded on to the map with only the first hour being visible. As the user clicks the button to advance through time, the appropriate tiles are made visible while the previous hour's tiles are removed from the map. The process of changing the visibility of layer is computationally minimal compared to the task of adding the tiles. When the number of map tile layers proceeding the currently visible layer reaches two, the next four hours of tiles are invisibly added to the map. This process ensures that a seamless flow of visual information is presented to the user.

4 Default Behavior vs. Real-time Bursts

This section presents two contrasting views of the POI-driven city pulse. First, the default behavior view aims at representing the constant and steady changes in activities conducted in the city. Temporally, the city goes through changes in activity dominance. This implies that specific activities, and the POI (types) at which these activities take place vary in intensity through out the day/week. This leads one to describe *Coffee Shops* as being mostly visited during the morning while *Nightclubs* are most active at night. This, we termed the Foursquare population's *Default Behavior* (towards POI).[13]

While humans are often described as creatures of habit (and the temporal bands support this), on an individual level, our behavior is often quite spontaneous and unpredictable. Analysis and visualization of these phenomena cannot be explored by looking at the POI ecosystem as a whole, but rather at a large scale or neighborhood level. It is virtually impossible to look at an overview of a city and attempt to understand the individual activities and behaviors of every inhabitant. Existing research by Cranshaw et al. [5], explored this phenomena of UGC-driven neighborhoods previously, but in a very different way. The authors show that a city can be split into subregions based on the social media contents generated by its residents. Our work takes a very different approach looking at real-time information presented in subregions rather than individual neighborhoods. For this reason, the *Social Burst View* view was developed.

[13] We are well aware of the fact that Foursquare is a biased data sources and thus our POI Pulse is biased (but this is not the focus of this paper).

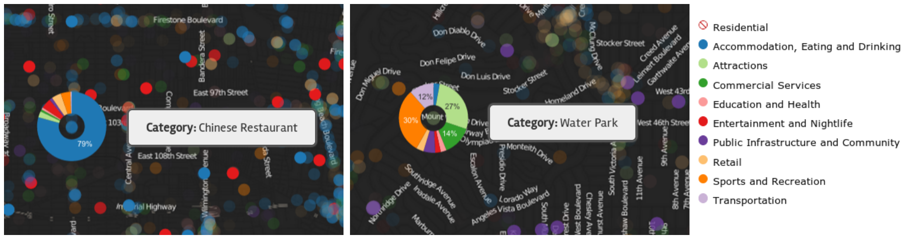
4.1 Default Behavior

From a systems architecture standpoint, the Default Behavior is accessible by zooming and panning through all map zoom levels. Visual exploration of default temporal behavior and spatial patterns is also encouraged by panning through time (clicking the *Next Hour* button) or jumping to a selected time (clicking the clock button and selecting from a drop down list of hours and days). This shows how the POI-driven urban system changes over time. At the initial map scale, a single color value is used to represent all user-contributed POI in the Greater Los Angeles area. Advancing through time while visualizing the POI in this way provides the user with a better understanding of the flow of the city as a whole. This view is essential to understanding which regions of the city are dynamic and the overall variability in activity level for the entire region. Increasing the map scale by one zoom-level, the user is presented with new upper-level classes. Again, panning through both space and time, the viewer gains a better understanding of the distinction between classes. As the opacity value of each POI marker changes, the user is made aware that the level of activity is changing both between and within classes. For example, the class of *Entertainment and Nightlife* is very prominent at 12am on Sunday while it is completely overshadowed by categories such as *Commercial Services* on Monday at 9am.

Zooming in further, the data format switches from image to vector tiles. While the color scheme, marker size, and opacity do not change between zoom levels, the capabilities of the vector data format allow for much greater user interaction. Hovering one's mouse over any POI between zoom levels 13 and 16 results in the Foursqure POI type name appearing beside the POI as well as a *donut-pie chart* surrounding the marker. The donut-pie chart is a technique employed to visually explain the OS class probabilities determined for each POI type in section 2.4 thus going beyond binary classification.

The value of being able to interact with the map through mouse events, for example, is that one can visually explore the probability distribution of classes for each individual POI. The standard marker visualization forces each POI to be assigned a single color representing a single class, but in actuality, the POI may exhibit high probabilities in more than one class and the primary marker color could be ascertained by a very small margin. When the user hovers over a POI, the donut-pie chart is displayed, demonstrating the multi-class characteristics of the venue. Each portion of the donut represents a category that contributes to this venue, and the color of each portion reflects the class. The size of each portion is defined by the percentage of this contribution based on the learned SVM model.

Figure 7 shows mouse-over interaction with two different POI. The central marker in Figure 7a is styled blue indicating that the primary OS class for this POI is *Accommodation, Eating and Drinking*. The accompanying donut-pie chart clearly shows that the highest probable classification for this POI **type** is indeed *Accommodation, Eating and Drinking* followed by small fractions of *Entertainment and Nightlife* and so forth. Comparatively, Figure 7b shows the prominent class for *Water Park* to be *Sports and Recreation* which makes sense given that users of the geosocial application are likely to visit a water park to engage in physical activities and recreation. The second highest class, as shown by the donut-pie char is *Attraction* and it is the second highest



(a) Chinese Restaurant (b) Water Park

Figure 7: Donut-pie charts showing OS category probabilities for two different POI.

probability by only a few percentage points. It is this ability to explore the discrepancies between classes, to dig into the underlying data, that is the true power of such an application.

4.2 Real-time Burst Mode

Understanding the pulse of a city involves not only looking at the city as a whole, but exploring the individual subregions or neighborhoods. What are people talking about in this part of the city? What places are popular right now? These are questions that should be asked, not at a city-wides scale, but rather at a local scale where the contents can be understood. Recent work by Purves et al. [14] explored this notion of describing place based on data contributed to geosocial applications such as Flickr and Geograph. In addition, the *LIVE Singapore!* project [10] allows individuals to access a range of real-time information from a variety of sources as well as contribute back to the system. While it does not include default temporal behavior as a foundation, it does offer real-time access to an assortment of city sensors.

Microblogging applications such as *Twitter* offer users the ability to geo-tag their content before publishing it. By accessing the streaming API,[14] these tweets can be added to the map immediately after they are published, providing the user with (near) real-time information on what is happening in a certain region. Additionally, Foursquare provides current check-in counts for any venue in their dataset through their rate-limited API.[15] This information is valuable in that it shows the true popularity of both Foursquare as a service, and the POI at which its users choose to check-in. Clicking the *Burst Mode* button immediately changes the temporal state of the map to the current hour of the day and week and begins to show real-time tweets and check-in counts based on the map view-port.

4.2.1 Real-time tweets

The Twitter streaming API offers users the ability to filter public streaming tweets by a specific geographic region. A listener service provides bounding box coordinates of

[14] https://dev.twitter.com/docs/streaming-apis
[15] https://developer.foursquare.com/docs

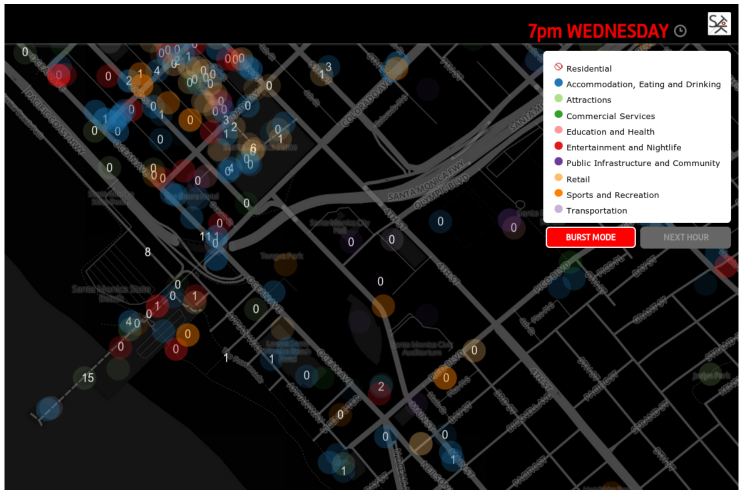


Figure 8: Real-time Check-in counts for Santa Monica at 7pm on a Wednesday

the study area and all tweets geotagged within the region are inserted in to a PostGIS database table. Individual tweets older than one minute are purged from the database complying with Twitter's terms of service.[16] Though twitter claims to restrict all tweets accessed through the filter streaming API to 1% of the real-time data, the influence of adding a geographic filter is not fully known. The average rate of tweets over a 24 hour period filtered to within the Greater Los Angeles area is approximately 113 per minute.

On the client/browser side, an asynchronous JavaScript (AJAX) request is made every 1000ms to a server side handler. The JavaScript request provides the view-port extent of the browser in geographic coordinates in order to restrict the returned tweets to only those within the user's view extent. In addition, only those tweets published within the last 2 minutes are requested. Upon return, the tweets are added to the map via a *D3* vector layer which produces an animation that mimics water droplets (Figure 9). The animation lasts for 1000ms while another request is made to the server.

Given the shear number of tweets, it is not technically prudent nor cognitively reasonable to display tweets on a small scale map. Recognizing this, users are given the option to view live tweets only within specific regions. The factor that determines the size and zoom scale of these regions is the number of POI within the view-port. A threshold of 1000 POI inside a view-port is the value at which users are given the option to view live tweets. Statistically, POI density is a good indication of neighborhood popularity, as the original POI were generated through crowd-sourcing means. It is important to note that this threshold is set independent of zoom level. As Figure 9 indicates, in some cases (Santa Monica Pier for example) the map scale will need to be quite large in order to fit less than 1000 POI in a view-port. Alternatively, parts of South-East Los Angeles reveal a lower POI density and therefore do not necessitate as large a map scale in order to visualize tweets.

[16] https://dev.twitter.com/terms/api-terms

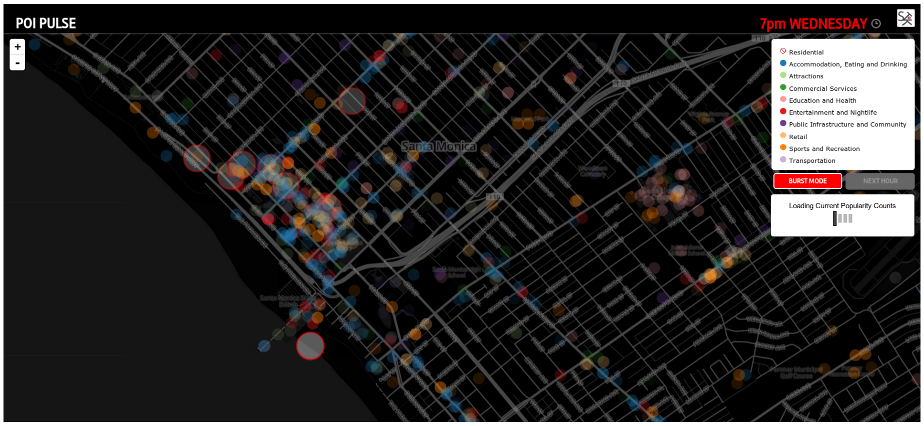


Figure 9: Tweets in Santa Monica (red circles with semi-transparent white fill)

4.2.2 Real-time POI popularity

While live geotagged-tweets offer insight into region-specific themes, the real-time geosocial popularity of POI in the region can also be determined. The Foursquare API permits requests to specific venues in order to determine the number of Foursquare users currently checked-in. Given API rate limits[17], a 500 POI view-port restriction ensures that any request made to a region returns only valid responses. From an system architecture perspective, when a user clicks the *Burst Mode* button, an AJAX request is made to the API which includes the geographic extent of the map view-port. The response from this request is then returned to the browser and check-in count values are added to the map for 500 POI, overlaid on top of the existing POI markers.

5 Conclusions and Future Work

Inspired by Foursquare's city pulse videos, 5 major shortcomings were identified that must be addressed to make the POI pulse useful from a scientific perspective and to contribute to the vision of information observatories for urban systems. Based on those shortcomings, we derived 4 theoretical and technical research questions that have to be successfully addressed to implement an improved urban pulse. In this work we addressed those questions by a combination of data-driven and theory-informed techniques to arrive at a semantics signatures-based POI taxonomy. We investigated how to seamlessly switch between multi-buffered image and vector tiles to implement a responsive Web portal that can handle over 170,000 POI (thus actively pushing the envelope of state of the art Web cartography). We studied the *tipping point* between those cached image and vector tiles, and finally proposed a method to seamlessly switch between a default mode of human behavior derived from empirical probabilities and streaming real-time geosocial data. We implemented the POI Pulse system as showcase

[17] API limitations require that a user request data through OAuth. Requests are limited to 500 per hour.

for our proposed solution.

In the future, we would like to add more services to the real-time burst mode. We are especially interested in a combination of platial, e.g., current check-ins, and spatial, e.g., Instagram pictures, data. As a proof-of-concept, we have discussed how to classify the human-generated content POI types to an administrative POI classification schema based on semantic signatures. However, we have only done so for level 1 and would like to add the second OS level in the future. Additionally, integration of POI and attribute data from alternative sources [13] would increase the variety and robustness of the proposed classification model. We also plan to add more bands, e.g., based on the place personalities proposed by Tanasescu et al. [17]. Next steps will also involve other methods of delineating class types, e.g., using a combination of color ramps to visually represent combinations of class probabilities and uncertainty.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] Michael Bostock and Jason Davies. Code as cartography. *The Cartographic Journal*, 50(2):129–135, 2013.

[3] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[5] Justin Cranshaw, Raz Schwartz, Jason I Hong, and Norman M Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *The 6th International Conference on Weblogs and Social Media*. AAAI, 2012.

[6] Sarah Elwood, Michael F Goodchild, and Daniel Sui. Prospects for VGI research and the emerging fourth paradigm. In *Crowdsourcing Geographic Knowledge*, pages 361–375. Springer, 2013.

[7] Mark Harrower and Cynthia A Brewer. Colorbrewer. org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[8] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. The fourth paradigm: data-intensive scientific discovery. 2009.

[9] Krzysztof Janowicz. Observation-driven geo-ontology engineering. *Transactions in GIS*, 16(3):351–374, 2012.

[10] Kristian Kloeckl, Oliver Senn, and Carlo Ratti. Enabling the real-time city: Live singapore! *Journal of Urban Technology*, 19(2):89–112, 2012.

[11] Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), 2013.

[12] Alan M MacEachren, Anuj Jaiswal, Anthony C Robinson, Scott Pezanowski, Alexander Savelyev, Prasenjit Mitra, Xiao Zhang, and Justine Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST 2011)*, pages 181–190. IEEE, 2011.

[13] Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. A weighted multi-attribute method for matching user-generated points of interest. *Cartography and Geographic Information Science*, 41(2):125–137, 2014.

[14] Ross Purves, Alistair Edwardes, and Jo Wood. Describing place through user generated content. *First Monday*, 16(9), 2011.

[15] Stuart Jonathan Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice hall Upper Saddle River, NJ, 2010.

[16] Ordnance Survey. Points of interest. http://www.ordnancesurvey.co.uk/business-and-government/products/points-of-interest.html, February 2014.

[17] Vlad Tanasescu, Christopher B Jones, Gualtiero Colombo, Martin J Chorley, Stuart M Allen, and Roger M Whitaker. The personality of venues: Places and the five-factors ('big five') model of personality. In *Computing for Geospatial Research and Application (COM. Geo), 2013 Fourth International Conference on*, pages 76–81. IEEE, 2013.

[18] Thanassis Tiropanis, Wendy Hall, Nigel Shadbolt, David De Roure, Noshir Contractor, and Jim Hendler. The web science observatory. *IEEE Intelligent Systems*, 28(2):100–104, 2013.

[19] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 520–528. ACM, 2011.